Cloud of Science. 2014. Т. 1. № 4
http://cloudofscience.ru
ISSN 2409-031X

Идентификация законов распределения по модельным и ограниченным выборкам данных в прикладной математической статистике

В. Б. Куликов

Нижегородский государственный технический университет им. Р.Е. Алексеева
603155, Нижний Новгород, ул. Минина, 24
e-mail: vb.kulikov@yandex.ru

Аннотация. Рассматривается метод идентификации плотности распределения вероятности случайных величин в стохастических системах, в том числе для модельных выборок ограниченного объема. Задача верификации решается для оценки предложенного метода восстановления полимодальных плотностей распределения. В качестве реконструируемой плотности принимается приближенное решение интегрального уравнения Фредгольма первого рода. Подтверждается вывод о необходимости увеличенного числа гладких функций для восстановления сложных (Коши, экспоненциальное) распределений.
Ключевые слова: Идентификация законов распределения, модельные выборки, уравнение Фредгольма первого рода, регуляризация.

1. Введение

В системах контроля сложными объектами, имеющих множество значимых внешних воздействий и обратных связей, важно не только корректно получать, но и интерпретировать измерительную информацию, экспериментальные данные. При анализе случайных процессов нельзя необоснованно пренебрегать эффектами меньшего по сравнению с основным порядком (локальными модами, малыми экстремумами, heavy tails-распределениями — Парето, Коши), так как в этом случае участки стохастических реализаций могут в ряде случаев порождать нежелательные трансформации управляющих сигналов.

Случайные помехи, погрешности измерений, несовершенство математических моделей и обработки данных способны менять вид распределения и приводить к некорректному применению алгоритмов (классический пример — фильтрация по Калману в системах управления). Поэтому проблема корректной, устойчивой и надежной идентификации и интерпретации характеристик и данных стохастических структур в прикладной математической статистике весьма актуальна.

В естествознании, квантовой механике, медицине, геофизике существует множество структур и процессов, при исследовании которых применяются сложные

методы идентификации законов, описывающих их стохастические или фрактальные характеристики. Особенностью таких характеристик является наличие сингулярных и многоэкстремальных (полимодальных) распределений. Например, в иммунологии идентифицируются распределения с числом мод 2–3, в теории автоподстраиваемых колебательных систем известны явления с 3–5 модами, в моделях, описываемых на основе решений уравнения Шредингера для квантовых систем, их может быть значительное число.

2. Исходная постановка задачи

Методы корректной идентификации законов распределения с указанной спецификой могут базироваться на решении обратных задач математической физики [1]. Обратные задачи, как правило, имеют некорректность в постановке: множественность «решения» и его чувствительность к погрешности исходных данных. Решение такого рода задач основано на методах регуляризации. Основу представленного подхода составляют методы корректного восстановления плотностей распределения случайных величин (СВ) и реализаций случайных процессов с апробацией методов на обширном фактическом материале в области клинической иммунологии. В данной сфере значительное число распределений показателей имеет специфические особенности в виде значительных уровней дисперсии, сложных законов распределения — многомодальных, негауссовых, негладкого типа — результат нелинейных эффектов в процессах образования/апоптоза клеток крови и лимфы, например в пролиферации Т-лимфоцитов [2].

Результаты научных школ в области создания устойчивых алгоритмов реализованы в виде программого обеспечения для приближенного решения интегрального уравнения Фредгольма I рода. Подынтегральная функция плотности вероятности является искомой величиной задачи. Правая часть уравнения соответствует эмпирической функции распределения для каждого показателя, например: уровня лейкоцитов, В-лимфоцитов, иммуноглобулинов, фагоцитарных чисел и других показателей крови и лимфы.

Используются ограничения на решения — непрерывность законов распределения изучаемых иммунологических показателей, их сосредоточенность на некотором отрезке (по диапазону изменения), гладкость формы плотности распределения.

С учетом последнего фактора, восстановление плотностей распределения всех иммунных показателей проведено на множестве тригонометрических функций с ограничением количества членов разложения N в зависимости от объема L наблюдаемых данных минимизацией гарантированного риска. Применение указанного подхода к обширному материалу иммунологических показателей позволило построить эмпирические законы распределения, классифицировать весь объем данных, и свести его к структурированной и строгой системе [3].

Предлагаемый подход принципиально отличается от классической схемы: не требует построения гистограмм, выбора числа и ширины интервалов группирования данных, не требует аппроксимации построенной гистограммы и

применения критериев согласия для проверки гипотез относительно закона распределения. Приближенное решение уравнения Фредгольма по выборкам малого объема позволяет корректно (на основе алгоритмов регуляризации) сразу построить кривую плотности распределения изучаемой СВ, минуя этап построения гистограмм (см. рис. 1–3).

3. Результаты идентификации законов распределения в иммунологии

В табл. для примера представлен фрагмент классифицированных данных, полученных в результате восстановленния эмпирической плотности распределения для пациентов-мужчин.

Таблица. Классификация законов распределения иммунологических показателей, восстановленных методом решения обратных задач (пациенты–мужчины)

	Классификационные признаки и характеристики						
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Иммуноло-гические показатели (мед. норма)	Объем выбор-ки, *L*	Класс распреде-ления	Степень поли-модаль-ности	Аппрокси мация норм. распреде-лением	Число членов разложе-ния для $p(x)$, N	Центр распре-деления	Вероят-ность попадания в интервал нормы
Лейкоциты, млн/л (4–9)	71	класс 7	1	да	4	8.70	0.65
Лимфоциты, % (19–37)	71	класс 1	2	нет	4	28.4	0.69
Лимфоциты, млн/л (0.7–3.8)	71	класс 7	1	да	4	2.43	0.91
Нейтрофилы п/я, % (1–5)	70	класс 1	2	нет	4	2.23	0.59
Нейтрофилы с/я, % (47–72)	71	класс 4	1	нет	3	59.4	0.69
Эозинофилы, % (1–5)	70	класс 5	1	нет	2	2.46	0.42
Моноциты, % (2–10)	71	класс 3	1	нет	2	7.45	0.45
Т-лимфоциты, % (40–90)	40	класс 2	3	нет	5	36.7	0.24
Т-лимфоциты, млн/л (0.5–3.0)	38	класс 7	1	нет	4	0.93	0.86
В-лимфоциты, % (2–30)	40	класс 4	1	нет	2	15.0	0.97
В-лимфоциты, млн/л (0.03–0.9)	38	класс 1	2	нет	5	0.40	0.97
Нулевые клетки, % (2–35)	40	класс 4	1	нет	2	48.1	0.06

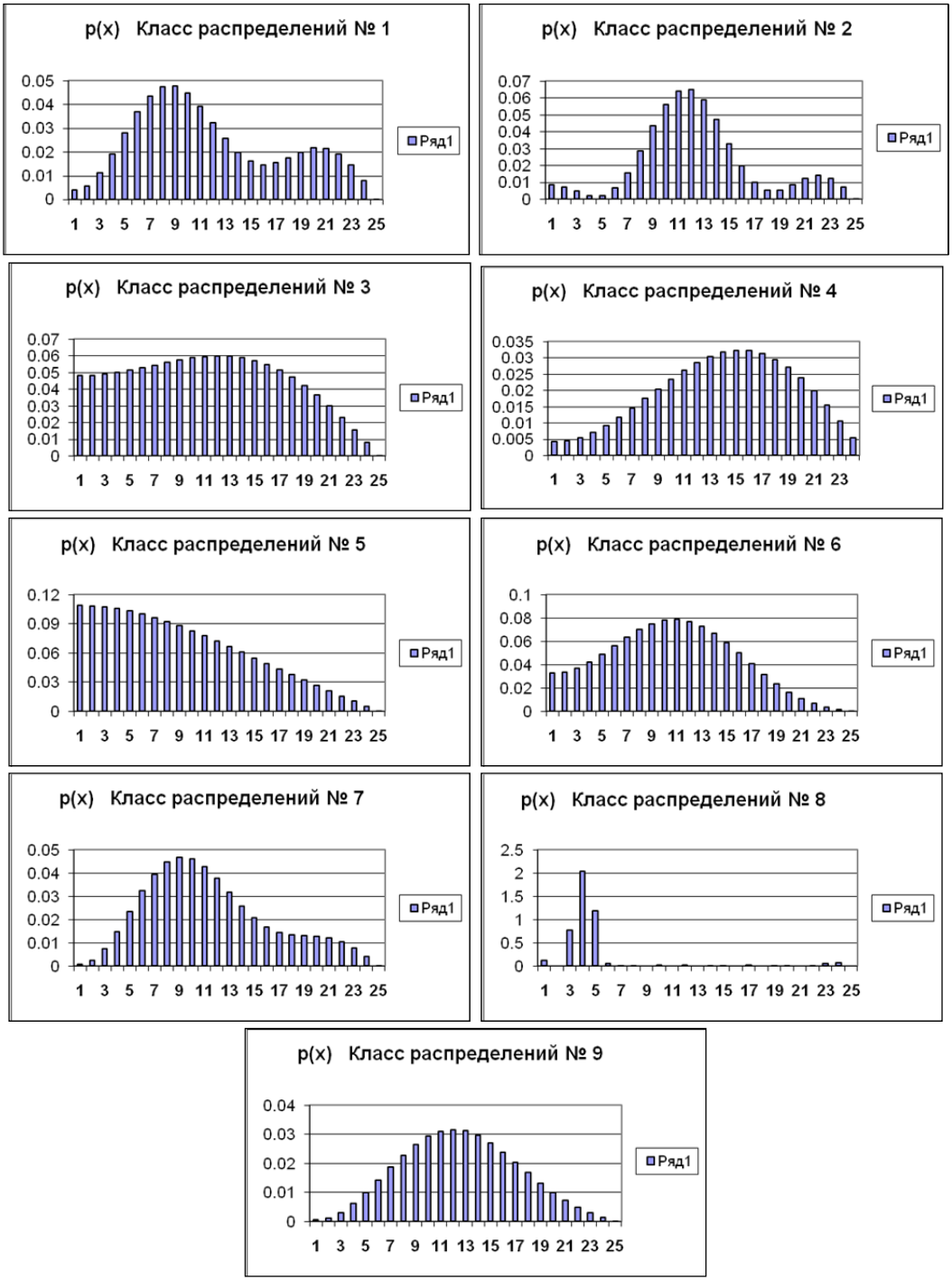


Рисунок 1. Графики плотностей p(x) иммунологических показателей пациентов после антибактериальной терапии, идентифицированные по выборкам малого объема и сгруппированные по девяти классам.

Рисунок 1 демонстрирует *типичные графики восстановленных плотностей распределения вероятности по классам иммунологических показателей*. Графики полученных функций содержат по 25 точек своих значений (можно 50–70 и т. д.) Рассмотрим их подробнее.

Из представленных плотностей распределения иммунологических показателей, только распределение № 9 имеет сходство с наиболее важным в теории вероятностей распределением — нормальным или близким к нему, при определенных параметрах, обобщенным распределением Рэлея. Последнее есть обобщение на случай — средние a_1 и a_2 не равны нулю — распределения на плоскости модуля вектора с координатами в виде независимых гауссовых случайных величин с одинаковыми дисперсиями. Указанное распределение имеет аналитическое выражение для плотности вероятности и находит широкое применение в статистической радиотехнике, радиофизике, механике сплошных сред, спектроскопии.

Обобщенное распределение Рэлея записывается таким образом:

$$p(x) = \frac{x}{\sigma^2}\exp\left\{-\frac{x^2+a^2}{2\sigma^2}\right\} I_0\left(\frac{xa}{\sigma^2}\right) \text{ при x > 0,}$$

$$p(x) = 0 \text{ при } x < 0,$$

где a — полярный радиус центра нормального распределения; $I_0(xa/\sigma^2)$ — функция Бесселя мнимого аргумента; σ^2 — дисперсия нормальных компонент вектора.

Распределение № 4 по своему виду сходно с логарифмически нормальным, зеркально отображенным (относительно оси ординат) распределением. Данный закон в исходном виде хорошо описывает поведение микро- и мезочастиц в биологии, медицине, статистической физике пластично деформируемых сплошных сред (например, поликристаллических соединений металлов) [4]. Логарифмически нормальное распределение в теории вероятностей также представляется в аналитической форме.

Иммунологические показатели, распределенные по классу № 3, представляют, по-видимому, предельную кривую плотности вероятности для класса № 4 при концентрации значений случайной величины в области нижней границы рассеяния.

Немногочисленный класс № 5 в первом приближении может быть описан любым односторонним распределением: треугольным (Симпсона), односторонним Гаусса (со значительной дисперсией), неравнобочным трапецеидальным распределением.

Шестой класс, за исключением области самых верхних значений, можно отнести к разряду распределений четвертого класса.

Есть также основание для хорошей аппроксимации его формульной зависимостью вида

$$p(x) = \frac{\alpha\,\lambda}{\pi}\left\{\frac{1}{\lambda^2 + (x_0 - x)^2} + \frac{1}{\lambda^2 + (x_0 - x)^2}\right\},$$

где α, λ — параметры распределения; x_0 — абсцисса наибольшего значения функции $p(x)$ — т. е. мода. В теории спектрального оценивания стационарных случайных процессов указанной функцией определяется спектральная плотность для процессов с корреляционной функцией вида:

$$k(\tau) = a \exp\{-\lambda\,|\,\tau\,|\}\cos(x_0\tau).$$

Распределение, отнесенное к восьмому классу, встретилось один раз («индекс нагрузки» — у женщин). Значения этого иммунологического показателя концентрируются в узкой области значений, а кривая распределения приближается по форме к дельта-функции. Можно предположить, что это нетипичная, квазидискретная форма распределения для иммунной системы — своего рода артефакт. Показатель «индекс нагрузки» — решение негладкого типа. Характерно, что два десятка членов не достигли аппроксимации на удовлетворительном уровне. Исследование возможностей аппроксимации для такого рода явлений в иммунологии и других биологических системах, а также в технике, технологических процессах и управлении представляет собой актуальную задачу.

Классы распределений 1, 2 и 7 в некотором смысле примыкают друг к другу. Два первых относятся к семейству кругловершинных многомодальных, причем первый ассиметричен, и его плотность вероятности имеет две различных моды, а второй — полимодален, с числом мод три и более. Моды высших порядков существенно ниже по своему уровню относительно главной моды. Следует отметить, что в этом классе обнаружено почти симметричное трехмодальное распределение (Т-лимфоциты, % — мужчины). В первом приближении плотности вероятностей для классов 1, 2 можно трактовать как суперпозицию гауссовых, соответствующим образом масштабированных и разнесенных по области изменения изучаемых показателей.

Одномодальное распределение класса 7 может интерпретироваться как непроявленная по второй, меньшей моде, форма двухмодального распределения класса 1.

Принцип «сложности» оценки плотности распределения: $N = N(L)$ — получил наглядное выражение в количестве требуемых для решения гармоник — минимум (3–7) для колоколообразных функций, в том числе, содержащих несколько локальных мод; максимум (20–25) — для компактно локализованных (малые уровни дисперсии).

Полученные численные значения восстановленных функций в дискретных точках (необходимого объема) используются для вычисления моментов любого порядка, а также энтропийных характеристик случайной величины, представляющей иммунологический параметр.

Для изучения функциональных состояний иммунной системы представляет интерес обнаружение многомодальных распределений у целого ряда показателей. В этом смысле, статистическое (стохастическое по своей природе) «поведение» части иммунных тел после интенсивной антибактериальной терапии можно сравнить, в частности, с многоорбитальным (по энергиям) распределением возбужденных электронов в теории лазерных эффектов, двухфотонно-возбуждаемых нелинейных процессов излучения люминесцентных спектров в генетических структурах [5]. Форма же распределения может ассоциироваться с волновыми функциями или распределением интенсивности освещенности для фраунгоферовой дифракционной картины, а также с другими фундаментальными физическими закономерностями.

В практическом плане сравнение известных оценок плотности вероятности иммунологических, гормональных и других показателей, исследование их трансформаций в состоянии здоровья и при терапевтических воздействиях, позволит вести мониторинг методов лечения, а также выявить глубинныt связи изучаемых явлений в клеточной биологии, микробиологии, клинической медицине с универсальными законами абиотического мира.

Кроме рассмотренного подхода исходный объем лабораторно-клинических данных — матрица размером примерно 80 на 30 — «пациенты-иммунные показатели» подвергался корреляционному анализу, подтвердившему многотаксонный характер полей корреляции и многомодальность целого ряда зависимостей.

Метод восстановления указанных законов распределения [3, 6] положительно зарекомендовал себя в плане вычислительной устойчивости, разрешающей способности по выделению и правильному позиционированию локальных особенностей. По сравнению с методом Парзена-Розенблатта он дает значительный структурный выигрыш, особенно на выборках малого объема.

4. Корректная идентификация распределений СВ на модельных выборках. Новые результаты для распределения Гаусса

Корректность метода, применяемого для восстановления плотностей распределения со многими модами необходимо дополнить соответствеющим анализом «искусственно созданных» распределений.

Для этого с использованием ряда результатов фундаментальной монографии [7] проведены исследования выборок, сгенерированных с помощью программ, разработанных авторским коллективом под руководством профессора Б. Ю. Лемешко и подчиняющихся различным законам распределения и их идентификация по предлагаемому подходу [8]. При этом объем выборок имел диапазон 100–300 единиц. Отдельно рассматривался случай смеси трех нормальных распределений при $L = 600$.

Изучались выборки случайных величин, относящиеся к теоретически абсолютно непрерывным распределениям: гамма-распределение, Коши, экспоненциальное, нормальное, Вейбулла. Исследовались выборки дискретных отсчетов «об-

разцовых» плотностей распределения, но фактически из рассмотрения исключались механизмы генерации сингулярных, свойственных теоретически фрактальным реализациям (Канторова «пыль» и др.).

Например, такие фактические данные можно получать по методу ЭГЭГ [9]. Они обладают, как показано автором, персистентностью, свойственной сингулярным (фракталоподобным) процессам. Как будет показано в следующих публикациях, применение регуляризирующих методик идентификации законов распределения в комплексе с алгоритмами предсказания позволяет существенно расширить также рамки анализа фракталоподобных структур естественной, техногенной, биологической, социальной и иной природы.

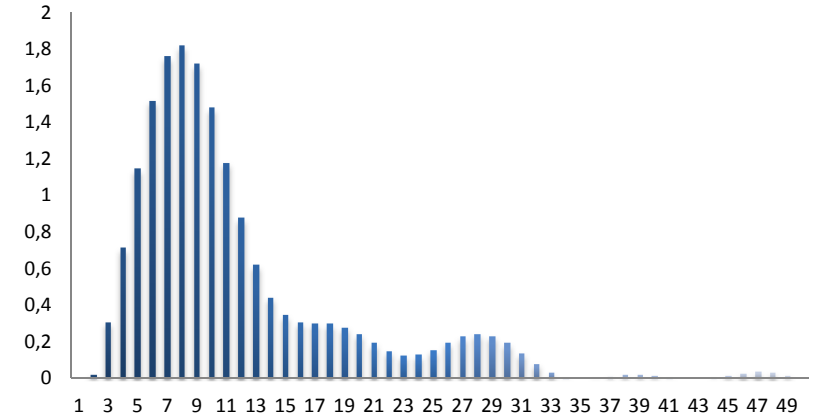
Результаты исследований показывают, что сохраняется вывод [6] о наличии особенностей сложных (Коши, экспоненциальное) распределений, что приводит к необходимости увеличенного числа гладких функций для представления приближенного решения: для Коши — примерно 20 гармоник, для экспоненциального 10–23. В последнем случае (рис. 2) форма кривой распределения структурно улучшается с увеличением объема выборки от 100 до 300 единиц.

Аналогичный вывод можно сделать и по нормальному закону (исчезает асимметрия и левый «хвост») с ростом объема выборки. Хотя в монографии [7] даже для L =100 критерии согласия дают хорошее соответствие принадлежности выборки к распределению Гаусса.

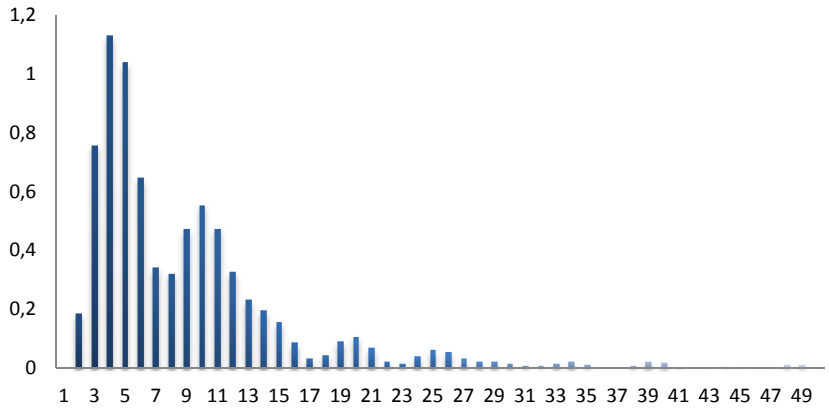
Предлагаемый метод принципиально выделяет главное отличие для выборок, сгенерированных по нормальному типу (исследовались объемы с L = 100, 200, 300, 600) — *число членов разложения N мало — и находится в диапазоне:* N = 2–4. При этом коэффициенты разложения решения по системе тригонометрических функций для выборки в 200 единиц имеют, к примеру, значение: $\alpha = 0.01393$; $\beta = -0.83330$; $\gamma = -0.50181$.

Автор полагает, что это фундаментальный признак распределений данного типа. Обнаруженное свойство случайных величин с распределением плотности вероятности Гаусса может быть использовано как дополнительный аргумент при идентификации данного закона распределения в прикладной математической статистике. «Колокольная» форма распределения Гаусса, правило «трех сигм» дополняются новым признаком идентификации.

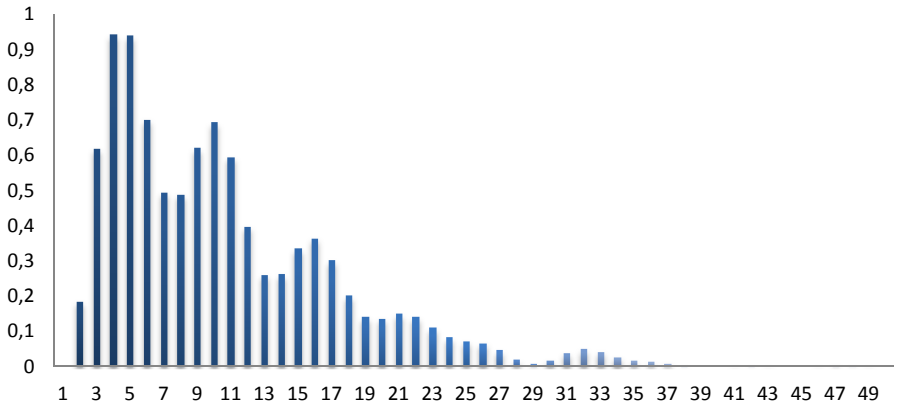
Для примера на рис. 2, 3 приводятся результаты восстановления плотностей теоретически абсолютно непрерывных распределений: экспоненциального и нормального. При этом объемы сгенерированных по [7] выборок последовательно увеличиваются (L = 100, 200, 300). Анализируется влияние на качество выборки алгоритма генерации показанных распределений и числа отсчетов. Показано, что имеет место корреляционная зависимость обоих факторов.



$L = 100$



$L = 200$



$L = 300$

Рисунок 2. Экспоненциальное распределение (после восстановления функции плотности вероятности; L — объем выборки)

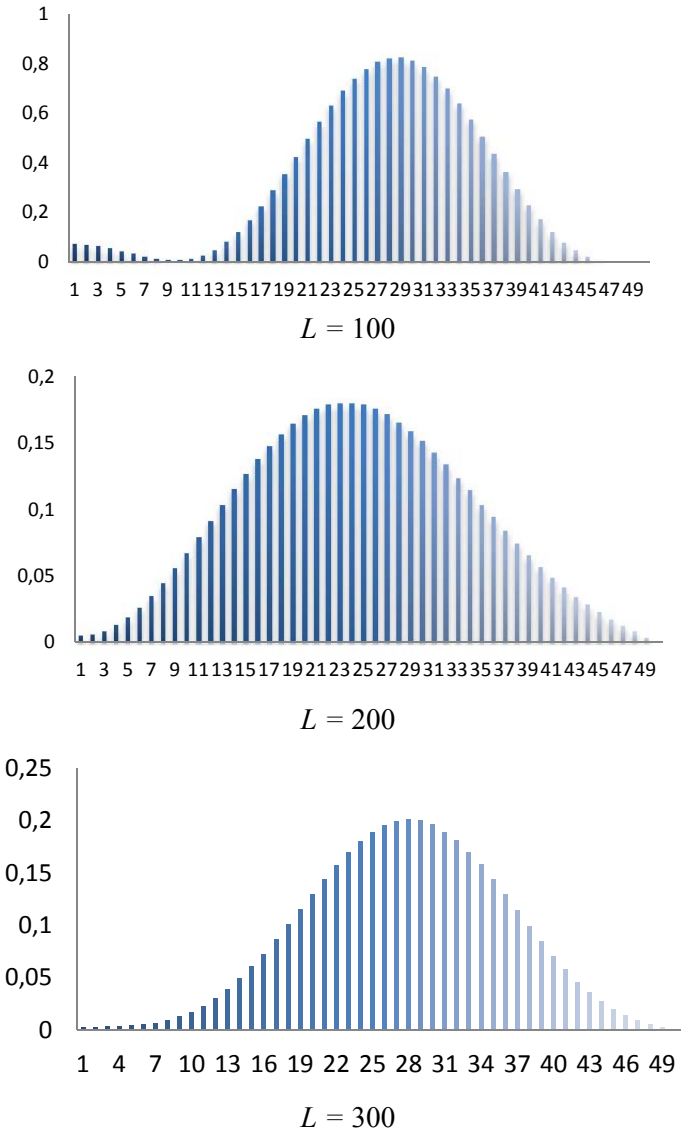


Рисунок 3. Нормальное распределение (после восстановления функции плотности вероятности; L – объем выборки)

5. Заключение

Выполненная идентификация эмпирических полимодальных распределений по выборкам малого объема изложенным способом позволяет считать его перспективным в области медицины, биологии, естествознания в целом. Исследования на модельных примерах показывают, что более простой, альтернативный метод Парзена-Розенблатта по разрешающей способности значительно уступает применяемому

подходу. Результаты восстановления плотностей моделируемых распределений: гамма-распределение, Коши, экспоненциальное, нормальное, Вейбулла с ограниченным объемом выборок прикладного значения демонстрируют высокие возможности метода; его универсальность и эффективность. Способность метода давать новые знания о «тонкой» природе стохастических явлений предлагается широко использовать в теории вероятностей и прикладной математической статистике.

Литература

[1] *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. — М. : Наука, 1986.

[2] *Бочаров Г. А., Лузянина Т. Б., Розе Дирк.* Математические технологии анализа пролиферации Т-лимфоцитов по данным проточной цитофлуориметрии // Российский иммунологический журнал. 2009. Т. 3(12). № 1. С. 13–22.

[3] *Куликов В. Б.* Восстановление полимодальных плотностей вероятности по экспериментальным данным в структурах со стохастическими свойствами // Вестник ННГУ им. Н. И. Лобачевского. 2014. № 1(1). С. 248–256.

[4] *Аратский Д. Б., Леонтьев Е. А. Морозов О. А., Солдатов Е. А., Фидельман В. Р.* Информационно-оптимальные методы в физике и обработке экспериментальных данных. — Н. Новгород: Изд. Нижегородского университета, 1992.

[5] *Агальцов А. М., Гаряев П. П., Горелик В. С., Рахматуллаев И. А., Щеглов В. А.* Двухфотонно-возбуждаемая люминесценция в генетических структурах // Квантовая электроника. 1996. Т. 23. № 2. С. 181–184.

[6] *Куликов В. Б.* Многомодальные законы распределения случайных величин в сложных стохастических системах и их идентификация // Материалы XIX МНТК «Информационные системы и технологии» ИСТ-2013. — Н. Новгород, 2013. С. 246–247.

[7] *Лемешко Б. Ю., Лемешко С. Б., Постовалов С. Н., Чимитова Е. В.* Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход. — Новосибирск: Изд. НГТУ, 2011.

[8] *Куликов В. Б.* Реконструкция функций плотности вероятности на модельных выборках методом регуляризации // Сборник трудов XXVII Междунар. науч. конф. «Математические методы в технике и технологиях» ММТТ-27 в 12 т. — Саратов, 2014. Т. 3. С. 169–172.

[9] *Нагорная М. Ю.* Применение фрактальных методов анализа к электрогастроэнтерографическим сигналам и их техническая реализация : автореф. дисс… к. т. н. — Самара : ПГУТИ, 2011.

Автор:

Куликов Владимир Борисович, аспирант Института радиоэлектроники и информационных технологий Нижегородского государственного технического университета им. Р. Е. Алексеева

Distribution Identification on Simulation Data and Limited Samples in Applied Mathematical Statistics

V. B. Kulikov

Nizhniy Novgorod State Technical University n.a. R.E. Alekseev
603155, 24, Minin st., Nizhniy Novgorod,
e-mail: vb.kulikov@yandex.ru

Abstract. Identified of the probability density function of model samples confinement. The task of verification is solved for the evaluation of the proposed method of restoring a multimodal distribution densities. As reconstructed density is taken approximate solution of integral Fredholm equation of the first kind. Is the regularization of the problem. Confirms the conclusion about the necessity of an increased number of smooth functions to recover the complex (Cauchy, exponential) distributions.
Keywords: Identification of distribution laws, model selection, the Fredholm equation of the first kind, regularization.

References

[1] *Tikhonov A. N., Arsenin V. J.* (1986) *Methods for solving ill-posed problems*. Moscow, Nauka. (In Rus)

[2] *Bocharov G. A., Luzyanina T. B., Rose Dirk*. (2009) Mathematical analysis techniques proliferation of T-lymphocytes according to the flow cytofluorimetry. *Russian journal of immunology*, 12(1), 13–22. (In Rus)

[3] *Kulikov V. B.* (2014) Restoring a multimodal probability densities on experimental data in structures with stochastic properties. *Journal of the Nizhny Novgorod University N. I. Lobachevsky*, 1(1), 248–256. (In Rus)

[4] *Aratscy D. B., Leontiev E .A. at al.* (1992) Information-optimal methods in physics and processing of experimental data. N. Novgorod. (In Rus)

[5] *Agaltsov A. M., Garyaev P. P. at al.* (1996) Two-photon-induced luminescence in genetic structures. *Quantum electronics*, 23(2), 181–184. (In Rus)

[6] *Kulikov V. B.* (2013) Multimodal distribution laws of random variables in complex stochastic systems and their identification. Proc. conf "Information systems and technologies" IST 2013. P. 246-247. (In Rus)

[7] *Lemeshko B. U. at al.* (2011) Statistical analysis of data, simulation and research of probabilistic regularities. Computer approach. Novosibirsk.

[8] *Kulikov V. B.* (2014) Reconstruction of probability density functions to model the samples by the method of regularization. Proc. intern. conf. "Mathematical methods in engineering and technologies" MMTT-27, Saratov. Vol. 3. P. 169–172.

[9] *Nagornova M. Y.* (2011) The use of fractal analysis methods to electrogastrographic signals and their technical implementation. Abstract of Thesis. Samara.