ИНФОРМАТИКА И УПРАВЛЕНИЕ В ТЕХНИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМАХ

УДК 519.213 ***DOI: 10.46960/1810-210X_2020_4_7***

В.Б. Куликов, А.Б. Куликов, В.П. Хранилов

ИДЕНТИФИКАЦИЯ И ВЕРИФИКАЦИЯ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ БИОМЕДИЦИНСКИХ ПОКАЗАТЕЛЕЙ НА ОСНОВЕ ОРТОГОНАЛЬНОГО БАЗИСА ЛЕЖАНДРА

Нижегородский государственный технический университет им. Р.Е. Алексеева

Предложена модификация базового метода идентификации полимодальных плотностей распределения случайных величин включением в алгоритм идентификации функционального базиса ортогональных полиномов Лежандра. Выполнена идентификация законов распределения группы иммунологических показателей в базисе Лежандра и верификация предложенного подхода. Модифицированный метод реализует дополнительную степень свободы при изучении характеристик иммунной системы по выборкам малого объема. Это позволяет в условиях дорогих или трудоемких анализов более обоснованно и надежно вести диагностику и интерпретировать результаты терапии. Применение функционального базиса Лежандра при исследовании функциональных состояний иммунной системы человека подтверждает наличие многомодальных (выявленных на тригонометрическом базисе) распределений у целого ряда показателей. Предложенный подход обладает лучшими возможностями при восстановлении равномерных или трапецеидальных распределений, а также класса уплощенных распределений типа шапо.

Ключевые слова: идентификация законов распределения, случайные величины, базис полиномов Лежандра, верификация, биомедицинские показатели.

Введение

В статье рассматриваются современные методы анализа экспериментальных данных медицинского мониторинга и экспертных систем для управления процессами лечения и диагностики. Основанные на алгоритмах решения обратных некорректных задач, данные методы включают в себя идентификацию законов распределения случайных биомедицинских показателей, верификационные аспекты тестовых задач. Они актуальны не только при терапии, но и при создании математических моделей структур, органов и систем человеческого организма, проявляющих стохастические свойства.

Идентификация искомых характеристик плотности вероятности выполнена методом решения интегрального уравнения Фредгольма I рода по ограниченой выборке. Сравниваются результаты, полученные при использовании как базиса Лежандра, так и основной базисной системы тригонометрических функций. В качестве медицинских показателей рассмотрены группы иммунологических данных для нескольких десятков пациентов по тридцати важнейшим показателям крови, лимфы, гормонов (Т, В-лимфоциты, лейкоциты, гранулоциты, антитела (иммуноглобулины), CD мембранные комплексы) после интенсивной антибактериальной терапии.

Постановка задачи

Основой подхода к вероятностному анализу биомедицинских данных (статистических по фактическому многообразию и стохастичеких по природе) служат методы корректной идентификации законов распределения случайных величин (СВ) [1]. Они базируются на решениях обратных некорректных задач, разработанных школой академика А.Н. Тихонова

[2]. Решение такого рода задач основано на методах регуляризации. В биомедицине значительное число распределений иммунных показателей имеет специфические особенности в виде высоких уровней дисперсии, сложных законов распределения (полимодальных, негауссовых) как результата проявления нелинейных эффектов в процессах, происходящих на клеточном и более высоком органно-системном уровне.

Ранее были опубликованы результаты системного подхода к корректному восстановлению плотностей распределения случайных величин и реализаций случайных процессов с апробацией методов на обширном фактическом материале в области клинической иммунологии и гастроэнтерологии [1,3].

Базовый метод идентификации плотностей распределения случайных величин опирается на принцип структурной минимизации риска и в конечном итоге заключается в решении эквивалентных систем линейных уравнений для нахождения коэффициентов разложения искомой плотности по выбранной системе базисных функций. В предположении гладкости формы плотности распределения восстановление законов 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. Тем не менее, оставался открытым вопрос о применении других функциональных базисов, отличных от традиционного тригонометрического. Такое исследование позволило бы, во-первых, верифицировать полученные ранее результаты [1], в которых восстановлены распределения со многими модами, во вторых – исследовать возможности новых базисов для различных законов распределения, включая близкие к равномерному. С этой целью рассмотрен базис на основе ортогональных полиномов Лежандра, приведены решения нескольких тестовых примеров, результаты идентификации и сравнения для обширной системы иммунологических показателей.

Тестирование функционального базиса на основе полиномов Лежандра

Напомним математические особенности полиномов Лежандра, используемые при модификации базового метода идентификации плотностей распределения. Полиномы Лежандра, ортогональные на отрезке [-1,1] с единичным весом $\rho(x) \equiv 1$, называются также сферическими полиномами. Формула Родрига дает общий член полинома Лежандра [4] (1):

$$P_n(x) = \frac{1}{n!2^n}\left[(x^2-1)^n\right]^{(n)}. \quad (1)$$

Базовый метод идентификации работает на интервале [0,1], потому необходим сдвиг базиса Лежандра с отрезка [-1,1] на рабочий интервал [0,1]. Используя линейную подстановку $x = \frac{2t}{T} - 1$, которая сдвигает область существования [-1,1] на промежуток [0,T], получим модифицированные (смещенные) ортонормированные полиномы Лежандра (2):

$$P_n(t,T) = \sqrt{\frac{(2n+1)}{T}} \cdot \frac{n!}{T^n} \cdot \sum_{k=0}^{n} \frac{C_n^k}{k!(n-k)!}(t-T)^k t^{n-k}. \quad (2)$$

При идентификации функция плотности предполагается непрерывной и сосредоточенной на отрезке [0,1], поэтому T = 1. Смещенные полиномы отличаются от классических тем, что первые содержат в каждом Pn слагаемые всех степеней, начиная с n-ой, а вторые – подразделяются на четные и нечетные.

Как и ранее, методы идентификации были реализованы в виде программого обеспечения для приближенного решения интегрального уравнения Фредгольма I рода. Подынтегральная функция плотности вероятности являлась искомой величиной задачи. Правая часть уравнения соответствовала эмпирической функции распределения для каждого показателя: уровня лейкоцитов, В-лимфоцитов, иммуноглобулинов, фагоцитарных чисел и других показателей крови и лимфы.

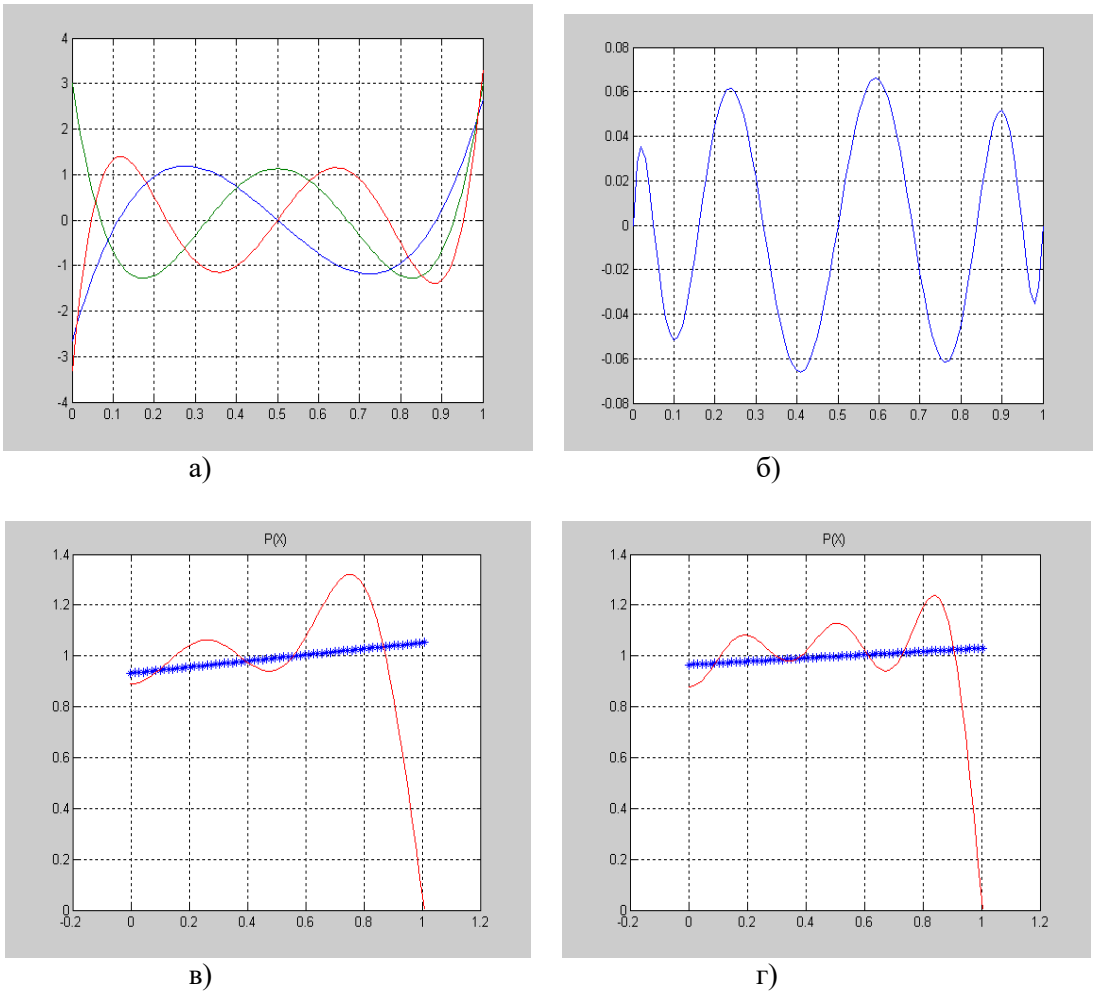


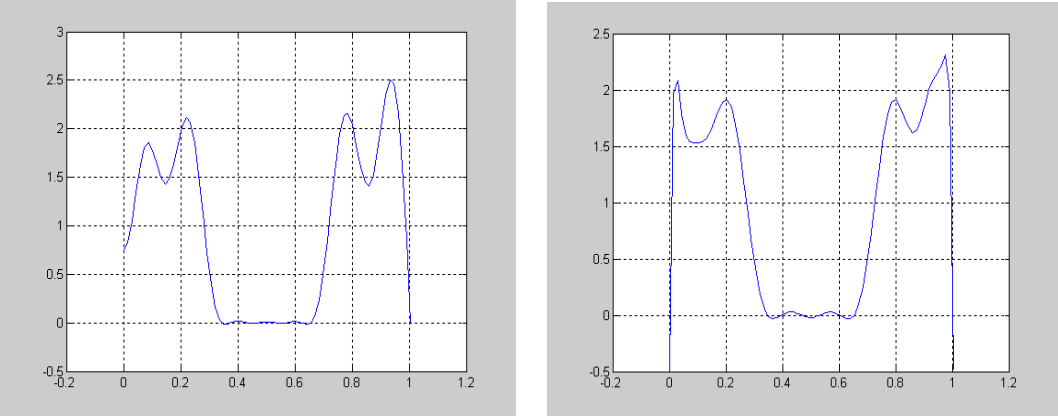
Рис. 1. Модифицированные (смещенные) полиномы Лежандра степеней n = 3, 4, 5
а) функция интеграла от полинома Лежандра 8 степени на отрезке [0,1]
б) графики идентифицированных плотностей равномерного распределения на отрезке [0,1] (объем выборок соответственно 1200 и 4800); жирная линия – базисные функции – полиномы Лежандра, тонкая – тригонометрический базис в), г)

На рис. 1(а) представлены модифицированные (смещенные) ортонормированные полиномы Лежандра степеней n = 3, 4, 5 на отрезке [0,1]. Как видно на изображении, система Лежандра должна обладать лучшими возможностями при восстановлении равномерных или трапецеидальных распределений, а также класса уплощенных распределений типа шапо [5]. Действительно, идентификация закона распределения тестовой равномерно распределенной выборки по системе тригонометрических функций существенно уступает по сравнению с вариантом ортонормированных полиномов Лежандра – рис. 1 (в, г). Для тригонометрического базиса (рис. 1 (в)) потребовалось четыре гармонических составляющих при колебательном характере графика $p(x)$, для базиса Лежандра – только два ортогональных полинома. При этом вычисленные по идентифицированным плотностям значения энтропийного коэффициента k равно 1,782 и 1,732 соответственно. Теоретическое значение для равномерного распределения k = 1,73.

Время идентификации с помощью тригонометрического базиса (рис. 1(г)) составляет 9 с, алгоритм с полиномами Лежандра требует 13 с. В соответствии с алгоритмом идентификации плотностей распределений на отрезке [0,1] требуется вычисление не только полиномов Лежандра, но и неопределенного интеграла для каждой степени полинома. Для повышения точности и ускорения работы вычислительного модуля авторами реализован аналитический алгоритм интегрирования. На рис. 1(б) показан график вычисленного интеграла от

полинома Лежандра 8 степени. Алгоритмы идентификации плотностей распределения по системе полиномов Лежандра реализованы в пакете MATLAB.

Второй тест и сравнение базисных систем представлены на рис. 2. Двухмодальная локально равномерная плотность распределения специальным образом формируется посредством комбинации функции *rand* пакета MATLAB (псевдослучайная последовательность двух прямоугольных импульсов). Объем выборки составляет 1 200 единиц. Случайная величина в некотором приближении моделирует множество Кантора первого порядка.



а) б)

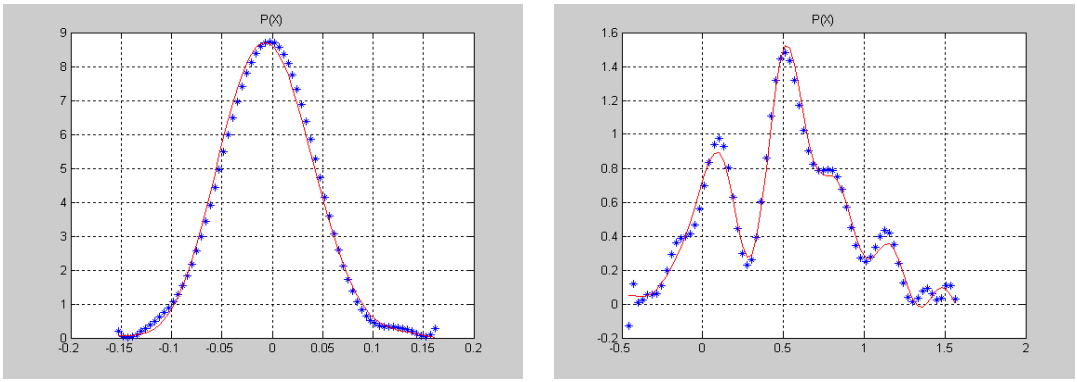
Рис. 2. Идентифицированные плотности двухмодального равномерного распределения на отрезке [0,1]. Объем выборки L = 1 200 отсчетов;
а) тригонометрический базис; b) полиномы Лежандра

Видно, что в базисе полиномов Лежандра (график *b*) можно получить более крутые фронты распределения и меньшие выбросы вершинных участков, чем в тригонометрическом базисе (график *a*). Вычислительные эксперименты показывают, что эта особенность резче проявляется в случае нескольких «импульсов» в законе распределения. Отметим, что оптимальное число членов разложения плотности распределения в обоих случаях примерно равно (15 и 17), но скорость идентификации по гармоникам выше примерно на порядок и составляет 1 с. Точность идентификации гладких распределений у обоих базисов примерно одинакова. На рис. 3 показаны восстановленные графики плотностей нормального (с малой дисперсией $\sigma = 0{,}05$) и распределения с тремя выраженными модами. Распределение (б) сформировано суммированием двух распределений: нормального (с малой дисперсией) и равномерного на отрезке [0,1].

Проведенные исследования на тестовых задачах верифицировали модифицированный метод на основе базиса Лежандра и показали новые возможности, включая восстановление негладких плотностей распределения и идентификацию распределений со многими модами. Отметим, что применение полиномов Якоби, содержащих несимметричные функциональные компоненты позволит наиболее «лаконичным» образом идентифицировать распределения Вейбулла, Пирсона, экспоненциального. В качестве базисных функций целесообразно также использовать полиномы Гегенбауэра, которые обобщают полиномы Лежандра и Чебышева и обладают возможностями адаптации к предполагаемому виду распределения [4].

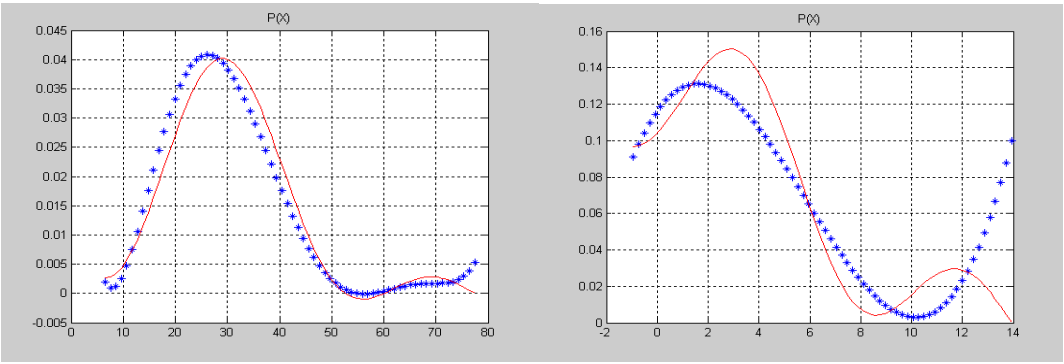
Исследование биомедицинских показателей на основе базиса полиномов Лежандра

Применение возможностей разработанного в [1] подхода к обширному материалу иммунологических показателей позволило восстановить и классифицировать весь объем данных, и свести его к структурированной и строгой системе. В настоящей работе сделана вери фикация результатов, полученных на основе системы тригонометрических функций.

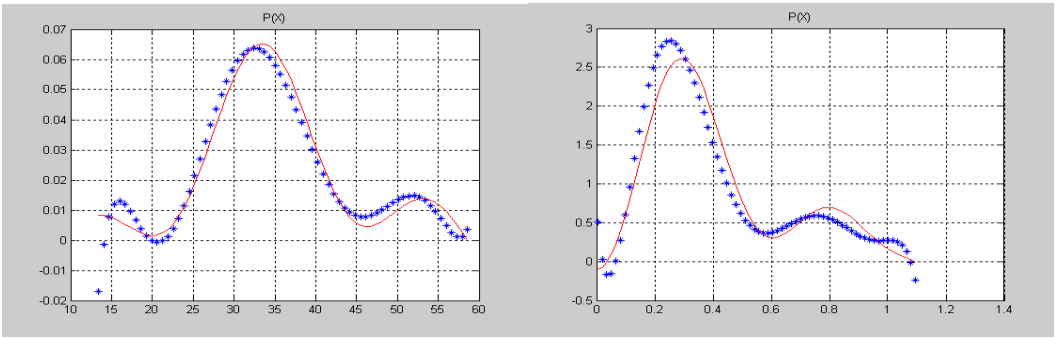


а) б)

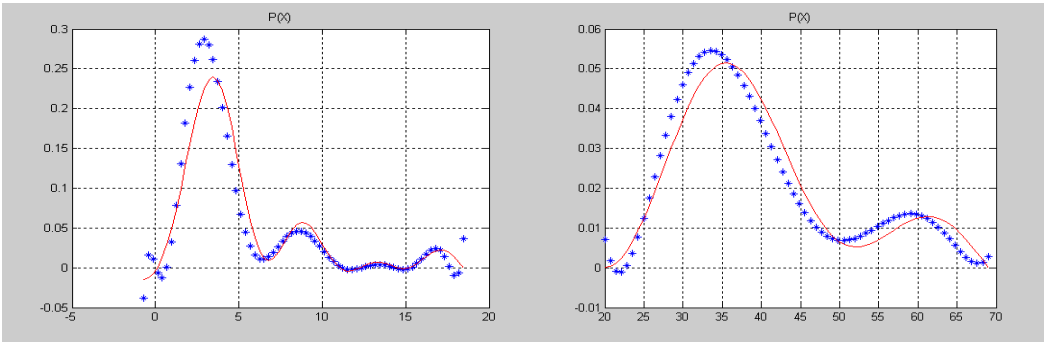
Рис. 3. Идентифицированные распределения с гладкими плотностями, L = 1200;
a) нормальное распределение; b) полимодальное распределение.
b) звездочка – базис нормированных полиномов Лежандра,
сплошная тонкая линия – тригонометрический базис



а) Лимфоциты, % (19-37) L =71 б) Нейтрофилы п/я, % (1-5) L = 70



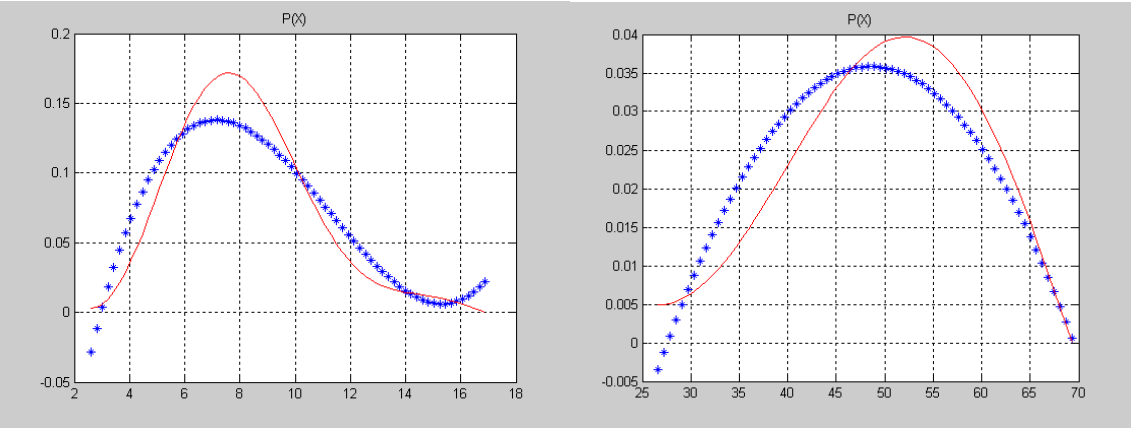
в) Т-лимфоциты, % (40-90) L = 40 г) В-лимфоциты, млн/л (0,03-0,9) L = 38



д) Тх/Тс, единиц (2,5-5,0) L = 40 е) Е-РОН-теофиллин, % (10-50) L = 40

Рис. 4. Идентифицированные полимодальные плотности распределения первой группы иммунологических показателей

Эти результаты дополнены идентифицированными законами распределения иммунологических показателей в базисе Лежандра. Важным обстоятельством является факт исследования выборок малого объема. Известно, что обработка данных такого объема является сложной методологической и математической задачей [6]. На рис. 4-6 представлены совмещенные графики идентифицированные плотности распределения трех групп иммунологических показателей пациентов-мужчин. Объем выборок L изменяется от 33 до 71.



а) Лейкоциты, млн/л (4-9) L =71 б) Нулевые клетки, % (2-35) L = 40

Рис. 5. Идентифицированные одномодальные плотности распределения второй группы иммунологических показателей

Время идентификации плотностей распределения: десятые доли секунды для тригонометрического базиса, не более 3 с – для базиса Лежандра.

На рис. 7 представлены результаты регрессионной оценки системы двух случайных величин у женщин: Y-лейкоцитов, млн/л и X-лимфоцитов, %. Методом экспресс обработки статистических данных построено поле рассеяния (корреляционное поле) и линеаризованный вариант линий регрессии. Наблюдается значительное рассеяние величины Y и протяженность величины X. Об этом свидетельствует и вычисленный коэффициент корреляции $\rho = - 0{,}39$.

Анализ корреляционного облака позволяет сделать следующие выводы:

1) наличие обособленной группы точек в интервале (36...44) по X (лимфоциты, %) свидетельствует о втором таксоне этого показателя, что подтверждает его двухмодальную форму распределения;

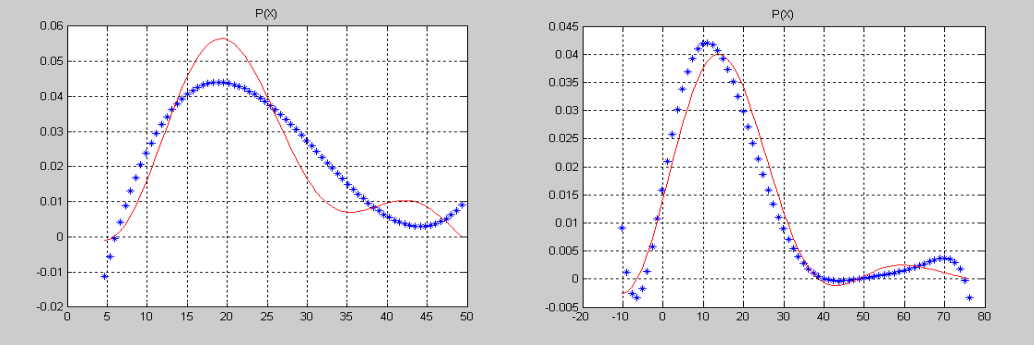
2) в силу значимости числа элементов этого подмножества ($\approx$ 25 %) данный факт нельзя отнести к промахам и случайности выборки;

3) точность метода восстановления плотности распределения по примененному принципу решения некорректно поставленных задач имеет наглядное доказательство;

4) методы регрессионного анализа для иммунологических показателей должны априорно задавать в большинстве случаев нелинейную модель и учитывать многотаксонный характер данных.

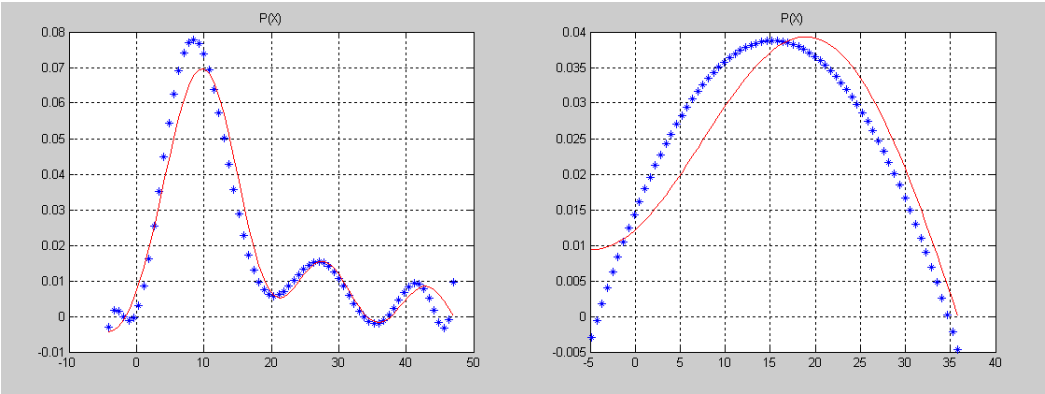
Рассмотрение аналогичных иммунологических показателей у пациентов-женщин в целом полностью подтверждает закономерность наличия высших мод в семействе лимфоцитарных показателей и CD-рецепторов.

Независимость этой выборки от мужской определяет дополнительные аргументы в обоснование предложенной модификации и классификации иммунологических данных.



а) CD 8+ (Т-киллеры), % (22,6 - 2,7)

б) CD 25 (Рецептор ИЛ2), % (14,0 - 3,5)



в) CD 71 (Рецептор транссферрина), % (9,4 - 1,8)

г) CD 26 (Активационный рецептор), % (13,6 – 5,0)

Рис. 6. Идентифицированные плотности распределения третьей группы иммунологических показателей с выборками минимального объема (L = 33 – объем выборки для каждого показателя)

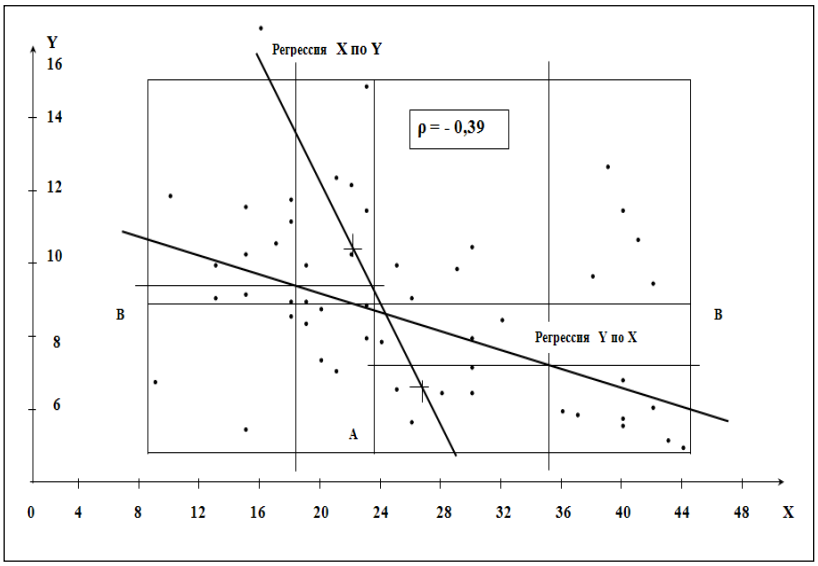


Рис. 7. Экспресс-метод статистической обработки иммунологических данных (пациенты – женщины): корреляционное поле и линии регрессии системы двух случайных величин (L = 52)

Выводы

Предлагаемый метод идентификации на основе функционального базиса Лежандра реализует дополнительную степень свободы при изучении характеристик иммунной системы по выборкам малого объема. Это позволяет в условиях дорогих или затратных по времени анализов, когда нет возможности увеличить объем выборки, более обоснованно и надежно вести диагностику и интерпретировать результаты терапии. Применение функционального базиса Лежандра при исследовании функциональных состояний иммунной системы подтверждает наличие многомодальных (выявленных на тригонометрическом базисе) распределений у целого ряда показателей.

Успешная идентификация эмпирических двух- и полимодальных распределений по выборкам малого объема предложенными методами позволяет считать восстановление плотностей вероятностей и сам подход адекватными проблеме точного оценивания.

Исследование на модельных примерах показывает, что альтернативный метод непараметрического восстановления Парзена-Розенблатта по эффективности и разрешающей способности значительно уступает применяемому подходу. Например, для относительно успешного восстановления функции трехмодального распределения требуется объем выборки в 1 200 отсчетов. Время вычислений по методу Парзена-Розенблатта составляет от 5 до 15 мин даже для одномодальной плотности [6]. На малых выборках (100-300 единиц) наблюдается пропуск значимых мод.

Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта № 19-07-00926_а.

Библиографический список

1. **Куликов, В.Б.** Восстановление полимодальных плотностей вероятности по экспериментальным данным в структурах со стохастическими свойствами // Вестник Нижегородского университета им. Н.И. Лобачевского. – 2014. – № 1(1). – С. 248-256.
2. **Тихонов, А.Н.** Методы решения некорректных задач / А.Н. Тихонов, В.Я. Арсенин. – М.: Наука, 1986. – 288 с.
3. **Kulikov, V.** The Identification of the Distribution Density in the Realization of Stochastic Processes by the Regularization Method / V. Kulikov // Appl. Mathem. Sciences. – Vol. 9. – № 137. – 2015. – P. 6827-6834.
4. **Дедус, Ф.Ф.** Классические ортогональные базисы в задачах аналитического описания и обработки информационных сигналов: учебное пособие / Ф.Ф. Дедус, Л.И. Куликова, А.Н. Панкратов, Р.К. Тетуев. – М.: Изд-во МГУ, 2004. – 141 с.
5. **Новицкий, П.В.** Оценка погрешностей результатов измерений / П.В. Новицкий, И.А. Зограф. – Л.: Энергоатомиздат, 1991. – 304 с.
6. Поршев С.В., Копосов А.С. Использование аппроксимации Розенблатта-Парзена для восстановления непрерывной случайной величины с ограниченным одномодальным законом распределения / С.В. Поршев, А.С. Копосов // Научный журнал КубГАУ. – 2013. – № 92(08). – С. 1-14.

Дата поступления
в редакцию: 29.09.2020

V.B. Kulikov, A.B. Kulikov, V.P. Khranilov

THE IDENTIFICATION AND VERIFICATION OF THE LAWS OF DISTRIBUTION OF BIOMEDICAL INDICATORS BASED ON THE LEGENDRE ORTHOGONAL BASIS

Nizhny Novgorod state technical university n.a. R.E. Alekseev

Purpose: Modern methods of analyzing experimental data from medical monitoring and expert systems for managing treatment and diagnostics processes are considered.
Design/methodology/approach: A modification of the basic method for identifying polymodal distribution densities of random variables is proposed by including the functional basis of orthogonal Legendre polynomials in the identification algorithm.
Findings: Identification of the laws of distribution of a group of immunological indicators in the Legendre basis, and verification of the proposed approach.
Research limitations/implications: The methods are based on algorithms for solving inverse ill-posed problems, include identification of the distribution laws of random biomedical indicators, verification aspects of test problems, and are relevant not only for therapy, but also for creating mathematical models of structures, organs, and systems of the human body that exhibit stochastic properties.
Originality/value: The method of identification based on the Legendre functional basis provides an additional degree of freedom when studying the characteristics of the immune system from small samples. This makes it possible to conduct diagnostics and interpret the results of therapy more reasonably and reliably in conditions of expensive or time-consuming analyses, when it is not possible to increase the sample size.

Key words: identification, distribution laws, random variables, Legendre polynomial basis, verification, biomedical indicators.