Метод наименьших квадратов с адаптивными весами при аппроксимации нелинейных функций в произвольном базисе

П. Г. Бурундуков[1], А. И. Каримов[2], А. А. Михайлов[3],
Т. А. Шпилевая[4], В. А. Пчёлко[5]

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

[1]burpavel2000@gmail.com, [2]aikarimov@etu.ru, [3]sasha.23443255@gmail.com, [4]chpiska324@mail.ru, [5]vapchyolko@stud.etu.ru

***Аннотация*. Метод наименьших квадратов широко используется для решения задач аппроксимации данных в различных областях, таких как обработка сигналов, статистический анализ и моделирование сложных систем. В работе исследуется применение метода наименьших квадратов и его модификаций для аппроксимации нелинейных функций с нормально распределённым шумом. Эксперимент проведён с использованием метода наименьших квадратов с адаптивными весами и набора базисных функций, позволяющих подбирать аппроксимирующую функцию для данных. Полученные результаты показывают, что аппроксимация функции на всем диапазоне данных, даже с использованием адаптивных весов, может приводить к игнорированию значимых точек данных или требовать расширения базиса, что не позволяет достичь требуемой точности. Предлагается использование локального подхода, основанного на разбиении функции на отдельные участки и применения МНК на них, для повышения точности аппроксимации. Полученные результаты свидетельствуют в пользу локального подхода.**

Ключевые слова: метод наименьших квадратов, нелинейные функции, нормально-распределённый шум, аппроксимация, базис, обработка данных

I. Введение

Метод наименьших квадратов (МНК) [1, 2] является одним из базовых методов для решения задач аппроксимации. Он применяется для нахождения оптимальных параметров модели путём минимизации отклонений между экспериментальными данными и значениями, предсказанными моделью. Основная цель метода заключается в минимизации суммы квадратов ошибок, что делает его мощным инструментом для анализа данных и построения моделей.

Широкое распространение МНК обусловлено его универсальностью и эффективностью. Этот метод легко реализуется, требует сравнительно небольших вычислительных ресурсов и находит применение в различных областях. В инженерной практике МНК используется для построения математических моделей на основе экспериментальных данных, например, для обработки облаков точек в горнодобывающей отрасли, где метод помогает устранять шум и восстанавливать 3D-модели [3]. Кроме того, МНК активно применяется в системах связи для детектирования сигналов в многоканальных системах MIMO, где он повышает производительность передачи данных [4].

Метод также применяется в анализе экспериментальных данных для краткосрочного прогнозирования, где его простота и универсальность делают его идеальным инструментом [5].

Традиционно МНК используется для анализа линейных систем. Примером такой модели является уравнение

$$y = \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n + \varepsilon, \qquad (1)$$

где y – зависимая переменная, $x_1, x_2, \ldots, x_n$ – независимые переменные, $\theta_1, \theta_2, \ldots, \theta_n$ – параметры модели, а ε – ошибка.

Несмотря на линейную природу метода, он успешно применяется и для ряда нелинейных задач за счёт преобразования нелинейных функций в линейные. Например, логарифмирование степенной функции $y = ax^b$ приводит к линейной зависимости $\ln(y) = \ln(a) + \mathrm{b} \cdot \ln(x)$, что позволяет применять МНК для оценки параметров. Подобные преобразования возможны для показательных и гиперболических функций, что значительно расширяет область применения метода.

Однако классический МНК имеет ряд ограничений. В частности, он предполагает равномерное влияние всех точек на аппроксимацию, что делает его чувствительным к выбросам и шуму в данных. Для решения этой проблемы разработаны модификации метода, среди которых особенно выделяется взвешенный метод наименьших квадратов (ВМНК) [6, 7]. Этот подход учитывает неодинаковую значимость точек, присваивая каждой точке вес в зависимости от её ошибки. ВМНК особенно полезен при наличии выбросов или неравномерного распределения ошибок, так как позволяет снизить влияние точек с большими отклонениями.

Например, в эконометрии и статистике МНК используется для анализа временных рядов, включая задачи прогнозирования и определения трендов, где часто применяют взвешенную версию метода для более точной обработки данных с различным уровнем шума [8].

Применение ВМНК требует построения весовой матрицы, которая корректирует вклад каждой точки в итоговое решение. Для нахождения параметров используется следующая формула:

$$\theta = (A^T W A)^{-1}(A^T W y), \quad (2)$$

где A – матрица базисных функций, y – вектор наблюдаемых данных, W – диагональная матрица весов.

Предлагаемый в данной работе локальный подход к аппроксимации данных на основе ВМНК предполагает разбиение области определения функции на интервалы, каждый из которых аппроксимируется независимо. Такой подход позволяет учитывать локальные особенности данных, однако требует сохранения гладкости модели на границах соседних интервалов, аналогично задаче поиска сплайнов. Для сохранения гладкости в оптимизационную функцию добавляется набор ограничений равенства производных на границах интервалов.

Одним из численных алгоритмов оптимизации, хорошо зарекомендовавших себя при решении задач с ограничениями, является метод внутренней точки. Данный метод учитывает, как равенства, так и неравенства в ограничениях [9].

В данной работе рассматривается применение ВМНК для аппроксимации нелинейных функций в условиях наличия нормально распределённого шума. Основное внимание уделено исследованию возможностей улучшения точности аппроксимации путём учёта локальных особенностей функций.

II. Методы

Для проведения исследования были выбраны несколько нелинейных функций, чей вид будет рассмотрен в следующем разделе. Для моделирования данных была использована функция с нормально распределённым шумом, задаваемая выражением:

$$y_i = f(x_i) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), \quad (3)$$

где $f(x_i)$ – это функция, описывающая зависимость данных от переменной x_i, а ε_i – случайная ошибка, распределённая по нормальному закону с нулевым математическим ожиданием и дисперсией σ^2.

Поскольку в реальных задачах точная структура аппроксимируемой функции часто неизвестна, был сформирован базис возможных функций для аппроксимации данных.

В качестве базисных функций использовались: $\theta_k(\mathrm{x}) = \{1, \mathrm{x}, \mathrm{x}^2, \sin(\mathrm{x}), \sin(2\mathrm{x}), e^{-0.5x}, e^{-x}\}$.

Данный набор функций позволяет учитывать, как полиномиальные, так и гармонические компоненты, обеспечивая достаточную гибкость модели. Выбор базиса обусловлен необходимостью сбалансировать сложность модели и её способность адекватно описывать локальные особенности данных.

Для определения первоначального веса точек использовался модифицированный алгоритм `dbscan` [10]. В отличие от базовой версии, где вес точек, попадающих в кластеры равен 1, а не попадающих – 0, в нашем подходе все точки участвуют в процессе определения весов. Этот метод кластеризации позволяет выявлять плотные области данных и определять центры масс кластеров. В процессе кластеризации, для каждой плотной области, вычисляется её центр как среднее значение координат всех точек, относящихся к этому кластеру. Вес каждой точки рассчитывался на основе её расстояния до ближайшего центра масс (d_i):

$$d_i = \min_k \sqrt{((x_i - C_{x_k})^2 + (y_i - C_{y_k})^2}, \quad (4)$$

где x_i, y_i – координаты рассматриваемой точки, C_{x_k}, C_{y_k} – координаты ближайшего центра масс.

Значение веса для каждой точки определялось по формуле $w_i = 1/(1 + d_i)$. Такой подход позволяет учитывать близость точки к плотным областям данных.

Для локального случая итоговый вес каждой точки ($w_{\text{итог}}$) определялся как комбинация глобального ($w_{\text{глоб}}$) и локального веса ($w_{\text{лок}}$) точки, что позволило более точно учесть её вклад в процесс аппроксимации. Вес точки x_i рассчитывался по следующей формуле:

$$w_{итог}(x_i) = \alpha w_{глоб}(x_i) + (1-\alpha) w_{лок}(x_i), \quad (5)$$

где $\alpha \in [0,1]$ – весовой коэффициент, регулирующий вклад глобальной и локальной составляющих в итоговую значимость.

Для сегментации данных на независимые участки, которые затем аппроксимировались отдельно, были наложены условия, обеспечивающие гладкость модели на границах сегментов. Эти условия включают:

$$f_k(x_i) = f_{k+1}(x_i), \; f_k'(x_i) = f_{k+1}'(x_i), \quad (6)$$

где x_i – граничная точка для двух соседних сегментов (последняя точка сегмента k, первая точка сегмента $k+1$), f_k и f_{k+1} – это аппроксимирующие функции, f_k' и f_{k+1}' – их производные соответственно. Эти ограничения обеспечивают гладкость модели на границах сегментов.

III. Экспериментальные данные

Для реализации и тестирования моделей аппроксимации использовалась среда MATLAB. На первом этапе для оценки весов точек данных применялась встроенная функция `dbscan`, что позволило эффективно кластеризовать данные, а затем вычислить веса, учитывающие близость точек к центрам кластеров. В частности, использовался стандартный алгоритм `dbscan` для выделения плотных областей.

Для нахождения аппроксимации с ограничениями была использована функция MATLAB `fmincon`, которая является мощным инструментом для решения задач нелинейной оптимизации с ограничениями. В качестве решающего был выбран описанный ранее метод внутренней точки.

A. Первая тестовая модель

В качестве тестовых данных для проверки предложенного подхода была использована функция:

$$y_{true}(x) = 2 \cdot sin(1.5x) + 0.5 \cdot e^{-0.3x}.$$

Эта функция была выбрана для тестирования, поскольку она сочетает гармоническую и экспоненциальную составляющие, что делает её достаточно сложной для точного описания с использованием простого базиса.

Моделирование проводилось на 500 точках в интервале $x \in [-5, 20]$. Итогами проведённого эксперимента являются два графика, представленных на рис. 1 и 2.

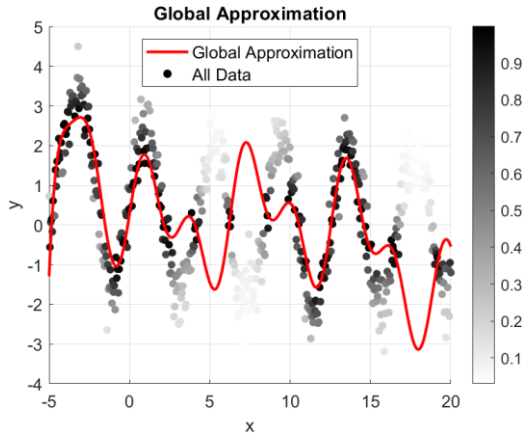

Рис. 1. Глобальная аппроксимация функции – не соответствие полученной аппроксимации представленным данным

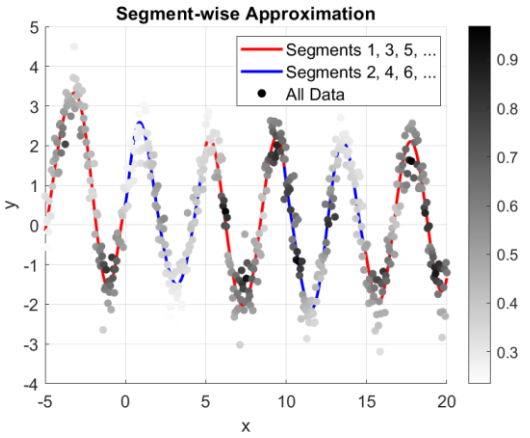

Рис. 2. Локальная аппроскимация функции – соответствие полученной аппроксимации представленным данным

Графики ошибок представлены на рис. 3.

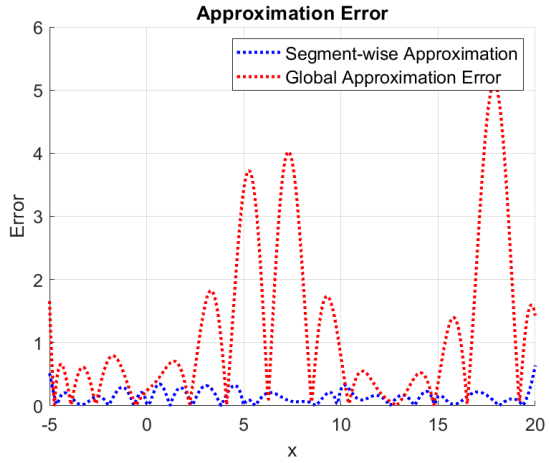

Рис. 3. Графики ошибок относительно истинной функции, красный – для глобальной аппроксимации, синий – для локальной

Результаты показали, что глобальная аппроксимация приводит к значительным погрешностям, особенно в областях с высокой кривизной функции, тогда как локальная аппроксимация позволяет более точно воспроизвести форму исходной функции, минимизируя ошибку аппроксимации.

Такая ошибка в глобальном случае связана с тем, что аппроксимирующая функция пытается минимизировать квадрат ошибки сразу на всём интервале значений. По итогу функция только частично попадает в 1-й пик, попадает во 2-й и 5-й, и не соответствует остальным.

B. Вторая тестовая модель

Вторая модель представлена функций вида:

$$y_{true}(x) = 0.8 \cdot sin(0.7x) + 0.5 \cdot \frac{log(1+5x^2)}{1+e^{-2x}} + 0.8 \cdot cos(1.5x^2).$$

Данная функция включает гармонические, логарифмические и экспоненциальные компоненты, что дополнительно усложняет задачу аппроксимации.

Моделирование проводилось на 500 точках в интервале x ∈ [–30, 30]. Итогами проведённого эксперимента являются два графика, представленные на рис. 4 и 5.

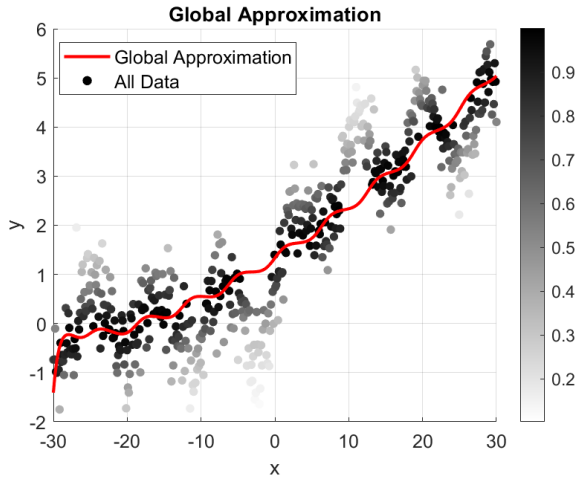

Рис. 4. Глобальная аппроксимация функции

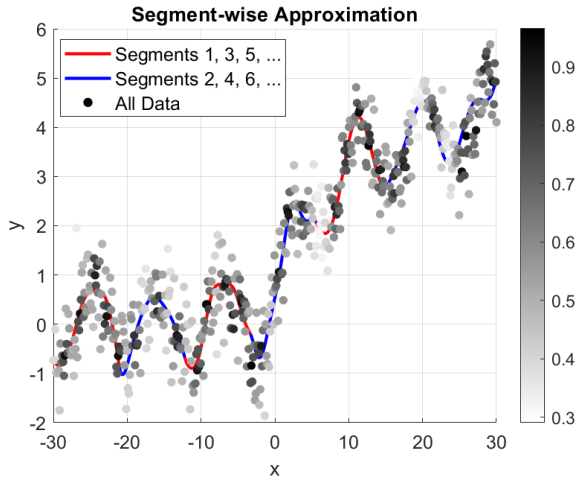

Рис. 5. Локальная аппроксимация функции

На графиках наблюдается соответствующая ранее полученным данным ситуация. В данном случае более наглядно видно, что, несмотря на попытку нелинейной аппроксимации в глобальном случае, полученный график сводится скорее к колеблющейся прямой линии, что соответствует методу наименьших квадратов.

График ошибок аппроксимации, соответствующий рис. 4 и 5, представлен на рис. 6.

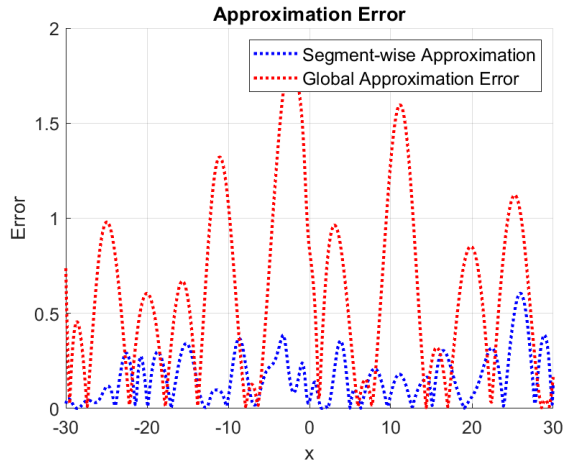

Рис. 6. Графики ошибок относительно истинной функции, красный – для глобальной аппроксимации, синий – для локальной

Из графиков 3 и 6 видно, что локальная аппроксимация функции работает более эффективно, а также подтверждают предположения о том, что для более успешной аппроксимации требуется локальный подход к учёту особенностей функции.

Таким образом, сравнение графиков 1 и 2, и в особенности 4 и 5 показывает, что глобальная аппроксимация функции приводит к значительному увеличению ошибки на участках с высокой кривизной. Визуально результат аппроксимации стремится к сглаженной прямой линии, не отражающей локальных особенностей. В отличие от этого, локальный подход позволяет учесть локальные особенности функции, обеспечивая плавные переходы между сегментами за счёт ограничения на равенство значений функции и её производной на границах.

IV. Заключение

Проведённое исследование демонстрирует, что применение метода наименьших квадратов с адаптивными весами в сочетании с локальной оптимизацией существенно повышает качество аппроксимации нелинейных функций, особенно в условиях наличия нормально распределённого шума. Применение модифицированного алгоритма кластеризации для расчёта начальных весов позволяет учитывать информационную ценность каждой точки данных, а комбинированный весовой коэффициент способствует более точному представлению локальных особенностей функции. Общая вычислительная сложность как глобального, так и локального подхода к аппроксимации определяется зависимостью $O(n_{sample} n_{features}^2)$, характерной для методов наименьших квадратов, использующих матричные операции, где n_{sample} – количество точек данных, а $n_{features}$ – количество независимых переменных. Однако благодаря сегментации функции на локальные интервалы, каждый из которых содержит меньшее количество точек n_{sample}, локальный подход позволяет сократить объём вычислений. Это снижение размера матриц в отдельных сегментах обеспечивает повышение общей вычислительной эффективности алгоритма.

Результаты экспериментов показывают, что глобальная аппроксимация зачастую приводит к сглаживанию важных характеристик функции и увеличению общей ошибки аппроксимации, в то время как локальный подход обеспечивает корректное моделирование даже для сложных нелинейных зависимостей. Использование метода внутренней точки для решения задачи оптимизации с нелинейными ограничениями гарантирует гладкость модели на границах сегментов и повышает устойчивость к выбросам.

Таким образом, разработанный метод представляет значительный интерес для практического применения в задачах анализа и моделирования сложных нелинейных данных, где требуется высокая точность и адаптивность модели при наличии шума. Дальнейшие исследования будут направлены на оптимизацию базиса функций на конкретных практических задачах и разработку динамических алгоритмов сегментации для ещё более эффективного учета локальных особенностей данных.

Список литературы

[1] Линник Ю.В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. 2-е изд. М., Физматгиз. 1962.

[2] Айвазян С.А. Прикладная статистика. Основы эконометрики. Том 2. М.: Юнити-Дана, 2001. 432 с.

[3] Jingli Wang, Huiyuan Zhang, Jingxiang Gao, Dong Xiao. Dust Removal from 3D Point Cloud Data in Mine Plane Areas Based on Orthogonal Total Least Squares Fitting and GA-TELM. Computational Intelligence and NeuroscienceVolume 2021, Issue 1. https://doi.org/10.1155/2021/9927982

[4] J. Wen, J. Li, H. Ge, Z. Zhou and W. Luo, "Orthogonal Least Squares Detector for Generalized Spatial Modulation," in IEEE Transactions on Wireless Communications, vol. 20, no. 8, pp. 5071-5082, Aug. 2021, doi: 10.1109/TWC.2021.3065383.

[5] Лыкова Е.А., & Лыкова Е.С. (2024). Применение метода наименьших квадратов для нахождения прогнозируемых значений. Современные инновации, системы и технологии - Modern Innovations, Systems and Technologies, 4(4), 0426–0435. https://doi.org/10.47813/2782-2818-2024-4-4-0213-0222

[6] Karimov A.I., Kopets E., Nepomuceno E.G., Butusov D. Integrate-and-differentiate approach to nonlinear system identification. // Mathematics, 9(23), 2021. 2999. https://doi.org/10.3390/math9232999

[7] Kariminejad, M.M., Sharifi, M.A. & Amiri-Simkooei, A.R. Tikhonov-regularized weighted total least squares formulation with applications to geodetic problems. // *Acta Geod Geophys* 57, 23–42 (2022). https://doi.org/10.1007/s40328-021-00365-1

[8] Прасолов А.В., Иванов Н.Г., Смирнов Н.В. Алгоритм оценки дисперсии для взвешенного метода наименьших квадратов // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2023. Т. 19. Вып. 4. С. 484-496. https://doi.org/10.21638/117011/spbu10.2023.405.

[9] Gill, P.E., Zhang, M. A projected-search interior-point method for nonlinearly constrained optimization. // *Comput Optim Appl* 88, 37–70 (2024). https://doi.org/10.1007/s10589-023-00549-1

[10] F K. Nafees Ahmed, T. Abdul Razak. An Overview of Various Improvements of DBSCAN Algorithm in Clustering Spatial Databases // International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016. Doi: 10.17148/IJARCCE.2016.5277