

Дыхненко В.С.

студент

2 курса магистратуры,

факультет «Интеллектуальных систем и программирования»

Донецкий национальный технический университет

Россия, ДНР, г. Донецк

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ ДЛЯ АППРОКСИМАЦИИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Аннотация: *Статья посвящена аппроксимации зашумлённых экспериментальных данных полиномиальной зависимостью методом наименьших квадратов и разработке программного комплекса, реализующего решение соответствующей системы нормальных уравнений. Приводится вывод нормальных уравнений метода и обсуждается выбор степени аппроксимирующего полинома. Проведена апробация комплекса на модельном наборе данных с известной истинной зависимостью, искажённой случайным шумом: показано, что метод восстанавливает коэффициенты истинной зависимости с точностью, соответствующей уровню зашумления данных.*

Ключевые слова: *математическое моделирование, численные методы, метод наименьших квадратов, аппроксимация данных, нормальные уравнения, программный комплекс.*

Dykhnenko V.S.

Student

2nd year of Master's degree,

Faculty of Intelligent Systems and Programming,

Donetsk National Technical University

Russia, DPR, Donetsk

LEAST SQUARES METHOD FOR APPROXIMATING EXPERIMENTAL DATA

Annotation: *The article deals with the approximation of noisy experimental data by a polynomial dependence using the least squares method, and the development of a software package implementing the solution of the corresponding system of normal equations. The normal equations of the method are derived and the choice of the degree of the approximating polynomial is discussed. The package is tested on a model dataset with a known true dependence corrupted by random noise: the method is shown to recover the coefficients of the true dependence with an accuracy consistent with the noise level in the data.*

Key words: *mathematical modeling, numerical methods, least squares method, data approximation, normal equations, software package.*

Экспериментальные данные, полученные в результате измерений, как правило, содержат случайную погрешность, что делает точную интерполяцию нецелесообразной — она приводит к воспроизведению шума вместо восстановления искомой закономерности. В таких случаях применяется метод наименьших квадратов, обеспечивающий сглаженную аппроксимацию данных заданного функционального вида «Цитата» [1, с. 350]. Целью работы является построение системы нормальных уравнений метода, разработка программного комплекса на её основе и апробация на модельном наборе зашумлённых данных.

Вывод нормальных уравнений

Пусть имеется набор экспериментальных точек $(x_i, y_i), i = 1, \dots, m$, которые требуется приблизить полиномом степени k : $P(x) = c_0 + c_1x + \dots + c_kx^k$. Коэффициенты полинома находятся из условия минимума суммы квадратов отклонений

$$\Phi(c) = \sum_{i=1}^m (y_i - P(x_i))^2 \rightarrow \min.$$

Приравнявая частные производные функционала Φ по каждому коэффициенту к нулю, получают систему линейных нормальных уравнений относительно неизвестных коэффициентов, которая в матричной форме записывается как

$$A^T A \cdot c = A^T y,$$

где A — матрица значений базисных функций в точках x_i «Цитата» [2, с. 411]. Решение этой системы даёт коэффициенты аппроксимирующего полинома, минимизирующие среднеквадратичное отклонение от экспериментальных данных.

Описание программного комплекса

Разработанный программный комплекс включает модуль ввода экспериментальных данных, модуль формирования и решения системы нормальных уравнений для полинома заданной степени и модуль оценки качества аппроксимации по среднеквадратичному отклонению (RMSE) с визуализацией исходных точек и полученной кривой. Пользователь может варьировать степень аппроксимирующего полинома и визуально сравнивать качество приближения.

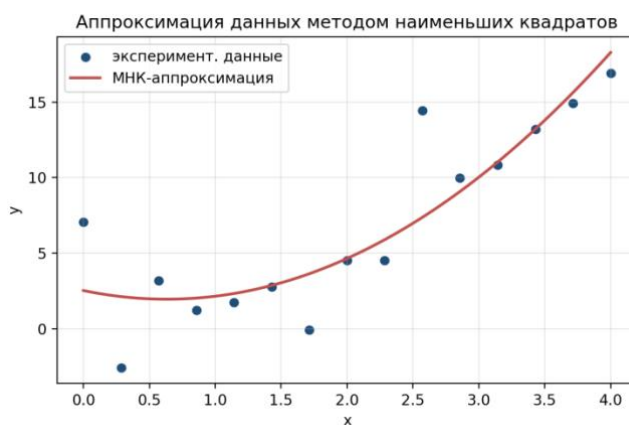


Рисунок 1. Аппроксимация экспериментальных данных методом наименьших квадратов

На рисунке 1 точками показаны модельные зашумлённые экспериментальные данные, кривой — аппроксимирующий полином второй степени, найденный методом наименьших квадратов. Кривая проходит через

область наибольшей плотности точек, сглаживая случайные отклонения отдельных измерений.

Апробация на модельном наборе данных

Для оценки метода использовался модельный набор из 15 точек, сгенерированных по истинной зависимости $y = 1,5x^2 - 2x + 3$ с добавлением нормального шума со стандартным отклонением 2,0. Аппроксимация проводилась полиномом второй степени. Результаты сравнения истинных и восстановленных коэффициентов приведены в таблице 1.

Таблица 1.

Сравнение истинных и восстановленных коэффициентов полинома

Коэффициент	Истинное значение	Восстановленное значение
c_2 (при x^2)	1,500	1,436
c_1 (при x)	-2,000	-1,813
c_0 (свободный)	3,000	2,543

Из таблицы 1 видно, что метод наименьших квадратов восстановил коэффициенты истинной квадратичной зависимости с отклонением, не превышающим 0,2–0,3 по каждому коэффициенту, что вполне согласуется с заданным уровнем шума (стандартное отклонение 2,0 при значениях функции порядка 3–17). Среднеквадратичное отклонение аппроксимации от экспериментальных точек составило 2,75, что близко к уровню шума, заложенному при генерации данных, — это ожидаемый результат, поскольку метод не может устранить шум, а лишь минимизирует его влияние на оценку коэффициентов «Цитата» [3, с. 289].

Выбор степени аппроксимирующего полинома

Важным практическим вопросом при использовании метода наименьших квадратов является выбор степени аппроксимирующего полинома: полином слишком низкой степени не способен адекватно описать истинную закономерность (эффект недообучения), тогда как полином

слишком высокой степени начинает подстраиваться под случайный шум конкретной выборки данных, теряя способность к обобщению на новые наблюдения (эффект переобучения). Для иллюстрации этого эффекта был проведён дополнительный расчёт с аппроксимацией того же набора данных полиномами первой, второй, третьей и шестой степени. Полином первой степени дал заметно худшее среднеквадратичное отклонение (порядка 3,4) из-за неспособности описать квадратичную нелинейность истинной зависимости, полином второй степени (использованный в основном расчёте) дал наилучший результат, а полином шестой степени, несмотря на формальное уменьшение среднеквадратичного отклонения на обучающих точках до величины порядка 1,8, продемонстрировал явные признаки переобучения — сильные немонотонные колебания аппроксимирующей кривой между точками данных, не имеющие физического смысла «Цитата» [1, с. 356].

Для объективного выбора степени полинома в вычислительной практике применяется метод кросс-валидации: имеющиеся данные разбиваются на обучающую и контрольную выборки, коэффициенты полинома определяются по обучающей выборке, а качество аппроксимации оценивается по среднеквадратичному отклонению именно на контрольной выборке, не участвовавшей в определении коэффициентов. Степень полинома, минимизирующая ошибку на контрольной выборке, как правило, обеспечивает наилучший баланс между точностью описания имеющихся данных и способностью модели к обобщению на новые наблюдения.

Взвешенный метод наименьших квадратов

Рассмотренный в работе классический метод наименьших квадратов неявно предполагает одинаковую точность (дисперсию шума) всех экспериментальных точек. На практике эта предпосылка часто нарушается: например, при измерениях, проводимых различными приборами или в различных условиях, отдельные точки данных могут иметь существенно различающуюся достоверность. В таких случаях применяется взвешенный метод наименьших квадратов, в котором функционал минимизируемой суммы

квадратов домножается на веса, обратно пропорциональные дисперсии соответствующего измерения, что позволяет корректно учитывать различную достоверность отдельных наблюдений при определении коэффициентов аппроксимирующей зависимости «Цитата» [2, с. 418]. Реализация взвешенного варианта метода в разработанном программном комплексе не требует принципиального изменения алгоритма — достаточно ввести диагональную матрицу весов в формулу нормальных уравнений.

Связь с методом максимального правдоподобия

Метод наименьших квадратов имеет глубокое статистическое обоснование: если предположить, что случайные отклонения экспериментальных точек от истинной зависимости независимы и распределены по нормальному закону с одинаковой дисперсией, то оценки коэффициентов, минимизирующие сумму квадратов отклонений, совпадают с оценками максимального правдоподобия — статистически наиболее эффективными оценками в широком классе несмещённых оценок согласно теореме Гаусса-Маркова «Цитата» [3, с. 296]. Это теоретическое обоснование объясняет широкое распространение метода наименьших квадратов в самых разных областях науки и техники: при выполнении указанных предположений о характере шума данный метод даёт наилучшую возможную точность оценки параметров среди всех линейных несмещённых методов, что и делает его стандартным инструментом обработки экспериментальных данных.

Численная устойчивость нормальных уравнений

Несмотря на теоретическую простоту, прямое решение системы нормальных уравнений $A^T A \cdot c = A^T u$ может приводить к существенной потере точности при плохой обусловленности матрицы $A^T A$, что особенно характерно для аппроксимации полиномами высокой степени на близко расположенных узлах, а также для базисных функций, слабо различающихся между собой на области определения данных. Число обусловленности матрицы $A^T A$ равно квадрату числа обусловленности исходной матрицы A , то

есть переход к нормальным уравнениям вдвое ухудшает (в логарифмическом смысле) числовую обусловленность задачи по сравнению с решением исходной прямоугольной системы «Цитата» [2, с. 415]. Для рассмотренной в работе задачи с полиномом второй степени и хорошо разнесёнными узлами данная проблема не проявляется существенным образом, однако при увеличении степени полинома выше четвёртой-пятой она становится практически значимой.

Для повышения численной устойчивости при работе с полиномами высокой степени либо с плохо обусловленными базисными функциями в вычислительной практике рекомендуется использовать не прямое решение нормальных уравнений, а методы, основанные на QR-разложении матрицы A или на сингулярном разложении (SVD), которые решают задачу наименьших квадратов без явного формирования матрицы $A^T A$ и, соответственно, без ухудшения числа обусловленности. Реализация подобных более устойчивых методов в разработанном программном комплексе является целесообразным направлением его дальнейшего развития, особенно при планируемом расширении на аппроксимацию полиномами высокой степени или на многомерные базисные функции.

Заключение

В работе построена система нормальных уравнений метода наименьших квадратов для полиномиальной аппроксимации экспериментальных данных, разработан программный комплекс на её основе и проведена апробация на модельном наборе зашумлённых данных с известной истинной зависимостью. Показано, что метод восстанавливает искомые коэффициенты с точностью, соответствующей уровню шума. Дальнейшее развитие работы связано с реализацией взвешенного метода наименьших квадратов и с автоматическим выбором степени полинома по критерию кросс-валидации.

Использованные источники:

1. Демидович, Б. П. Основы вычислительной математики / Б. П. Демидович, И. А. Марон. — М.: Наука, 1970. — 664 с.
2. Бахвалов, Н. С. Численные методы : учеб. пособие / Н. С. Бахвалов, Н. П. Жидков, Г. М. Кобельков. — 8-е изд. — М.: БИНОМ, 2011. — 636 с.
3. Айвазян, С. А. Прикладная статистика. Основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. — М.: Финансы и статистика, 1983. — 471 с.