УДК 519.245
MSC 65C05

Вестник СПбГУ. Математика. Механика. Астрономия. 2022. Т. 9 (67). Вып. 1

О выборе базисных функций регрессии и машинном обучении

С. М. Ермаков[1], *С. Н. Леора*[2]

[1] Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9
[2] Санкт-Петербургский государственный экономический университет,
Российская Федерация, 191023, Санкт-Петербург, наб. канала Грибоедова, 30–32

Для цитирования: *Ермаков С. М., Леора С. Н.* О выборе базисных функций регрессии и машинном обучении // Вестник Санкт-Петербургского университета. Математика. Механика. Астрономия. 2022. Т. 9 (67). Вып. 1. С. 11–22.
https://doi.org/10.21638/spbu01.2022.102

Как известно, в задачах машинного обучения широко используются средства регрессионного анализа, которые позволяют устанавливать связь между наблюдаемыми переменными и компактно хранить информацию. Наиболее распространенным является случай, когда функция регрессии описывается линейной комбинацией некоторых заданных функций $f_j(X)$, $j = 1, \ldots, m$, $X \in D \subset R^s$. Если наблюдаемые данные содержат случайную ошибку, то восстановленная по наблюдениям функция регрессии содержит случайную ошибку и систематическую ошибку, зависящую от выбора функций f_j. В данной работе указана возможность оптимального, в смысле заданной функциональной метрики, выбора f_j, если известно, что истинная зависимость подчиняется некоторому функциональному уравнению. В ряде случаев (правильная сетка, $s \leq 2$) близкие результаты могут быть получены с помощью техники анализа случайных процессов. Численные примеры, приведенные в данной работе, иллюстрируют существенно более широкие возможности предполагаемого подхода к задачам регрессии.

Ключевые слова: регрессионный анализ, аппроксимация, базисные функции, операторный метод, машинное обучение.

1. Введение. Пусть функция $f(X)$, $X \in D \subset R^s$, в каждой точке области D может быть вычислена или измерена с помощью некоторого прибора. Во многих случаях представляет интерес следующая задача. В предположении того, что $f(X)$ принадлежит некоторому линейному нормированному пространству F, требуется приблизить ее с помощью обобщенного многочлена $P_m(X) = \sum_{j=1}^m c_j f_j(X)$, где c_j — константы, а f_j — заданные функции из F. Если $f(X)$ измеряется с ошибкой в заданных точках $X_1, \ldots, X_n$, $n \geq m$, то при заданных $f_j(X)$ имеем задачу линейной по параметрам регрессии:

$$y(X_i) = \sum_{j=1}^m \hat{c}_j f_j(X_i), \quad y(X_i) = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

где ε_i — случайная ошибка, $\hat{c}_j$ — константы.

Наиболее изученным является случай, когда функция

$$f(X) = \sum_{j=1}^{m} c_j f_j(X) \tag{1}$$

известна с точностью до параметров c_j, и $\mathsf{E}\varepsilon_i = 0$,

$$\mathsf{E}\varepsilon_i \cdot \varepsilon_k = \begin{cases} \sigma_i^2, & i = k, \\ 0, & i \neq k, \end{cases}$$

σ_i^2 — константы, известные или определяемые в процессе эксперимента.

Задача нахождения (оценки) констант c_j представляет собой параметрическую задачу математической статистики. Наиболее распространенным методом ее решения является метод наименьших квадратов, с помощью которого параметры c_j определяются из условия

$$\sum_{i=1}^{n} \left(y(X_i) - \sum_{j=1}^{m} \hat{c}_j f_j(X_i) \right)^2 \cdot (1/\sigma_i^2) = \min,$$

$\hat{c}_j$ — это случайные величины такие, что $\mathsf{E}\hat{c}_j = c_j$ и дисперсия их минимальна [1].

Выбор функций $f_j(X)$, называемых базисными функциями регрессии, осуществляется с учетом сведений о функции f.

В действительности условие (1) обычно выполняется приближенно. Существует систематическая погрешность, которую нужно сделать минимальной. Если $f(X)$ — достаточно гладкая функция, то часто считают, что для малых областей D приближение с помощью алгебраического многочлена может быть приемлемым, c_j являются его коэффициентами.

Ниже мы рассмотрим новый (операторный) метод выбора базисных функций f_j и проиллюстрируем некоторые его преимущества и особенности.

2. Операторный метод. Суть операторного метода проще всего объяснить следующим образом. Если подлежащая аппроксимации функция удовлетворяет условию $Af = 0$, то в качестве базисных функций f_j можно выбрать собственные функции оператора A или близкого к нему в операторной норме оператора $\hat{A}$. Если оператор A самосопряженный, то известны экстремальные свойства таких приближений. В случае несамосопряженного оператора следует использовать разложение Фишера, на котором мы далее остановимся подробнее.

Пусть f принадлежит подпространству $\mathbb{F}$ гильбертова пространства и A — несамосопряженный оператор такой, что $Af \in \mathbb{F}$. Размерностью $r(A)$ оператора A называют размерность подпространства Af. Обозначим через H оператор AA^T, через φ_j $(j = 1, \ldots, r(A))$ — ортонормированную систему собственных функций оператора H. Имеет место представление [2, с. 47]

$$A = \sum_{j=1}^{r(A)} s_j(A)(\cdot\,, \varphi_j) \bigcup \varphi_j,$$

где s_j — соответствующие собственные числа оператора $H^{\frac{1}{2}}$, а $\bigcup H$ — полярное представление A, $\bigcup \varphi_j = \psi_j$.

Далее, если

$$A_m = \sum_{j=1}^{m} s_j(A)(\cdot\,, \varphi_j)\psi_j, \quad m \le r(A),$$

то среди всех m-мерных операторов A оператор A_m доставляет наименьшее значение $\|A - A_m\|$ равномерной норме разности $A - A_m$.

Если $K = I - A$ оператор такой, что верно равенство $Kf = 0$, и мы заменим в этом равенстве A его аппроксимацией A_m, то получим приближение f в виде

$$\hat{f} = \sum_{k=1}^{m} s_k(f, \varphi_k)\psi_k. \tag{2}$$

Аналогично поступаем, если оператор K является приближением к $I - A$.

Описанный алгоритм может быть частью процедуры машинного обучения [3]. Идея обучения состоит в следующем. Пусть $K(\Theta)$ — параметрическое семейство операторов. Используя аналитические выражения для функции f или ее численные значения, найдем оператор $K_0 = K(\theta_0)$, $\theta_0 = \operatorname{argmin}_{\theta\in\Theta}\|K(\theta)f\|$, если функция f задана без ошибки, или $\theta_0 = \operatorname{argmin}_{\theta\in\Theta}\mathsf{E}\|K(\theta)f\|$, если имеется случайная погрешность. Полагая $A = I - K_0$, найдем соответствующие φ_k и ψ_k и построим для некоторого подходящего m приближение $\hat{f}$. Для обучения можно рассмотреть несколько параметрических семейств операторов и несколько значений m.

Таким образом, описанная методика может применяться для выбора оптимального полинома $\sum_{j=1}^{m} c_j f_j(X)$ в задачах аппроксимации и регрессии.

Рассмотрим простейший пример. Пусть $\mathbb{F}$ есть подпространство дважды дифференцируемых функций на промежутке (a, b) и при $m = 2$ оператор A равен

$$A = \frac{\mathrm{d}^2}{\mathrm{d}x^2} + p\frac{\mathrm{d}}{\mathrm{d}x} + q. \tag{3}$$

Считаем функцию f заданной аналитически, а оценку параметров p и q произведем с помощью метода наименьших квадратов. Тогда p и q найдем из условия

$$\int_a^b (f'' + pf' + qf)^2 = \min.$$

Нужно решить систему

$$p\int_a^b f'(x)f'(x)\mathrm{d}x + q\int_a^b f(x)f'(x)\mathrm{d}x = -\int_a^b f'(x)f''(x)\mathrm{d}x,$$

$$p\int_a^b f(x)f'(x)\mathrm{d}x + q\int_a^b f^2(x)\mathrm{d}x = -\int_a^b f(x)f''(x)\mathrm{d}x.$$

Составим характеристическое уравнение

$$k^2 + pk + q = 0.$$

Если его корни k_1 и k_2 различны (мы не рассматриваем вырожденный случай), то $f_1(x) = \exp(k_1 x)$ и $f_2(x) = \exp(k_2 x)$.

Оператор (3), однако, не является самосопряженным (первая производная меняет знак). И в этом случае мы имеем

$$H = \frac{\mathrm{d}^{(4)}}{\mathrm{d}x^4} + (2q - p^2)\frac{\mathrm{d}^{(2)}}{\mathrm{d}x^2} + q^2.$$

Впрочем, из выполненных численных примеров следует, что использование только собственных функций оператора A даже в несамосопряженном случае обеспечивает значительное преимущество по сравнению с полиномиальной регрессией.

Не представляет труда провести обобщение приведенных примеров на случай дифференциальных операторов с постоянными коэффициентами и других более высоких порядков.

Наличие случайной ошибки у f не позволяет, вообще говоря, использовать дифференциальный оператор. В этом случае самым простым является применение разностного оператора. Такая методика рассматривалась в многочисленных работах по анализу временных рядов [4], где измерения $f(t)$ проводились в равноотстоящих по времени точках с шагом Δt. Результаты этих работ, с одной стороны, выходят за рамки регрессионной постановки, но, с другой, не позволяют рассматривать случай для неравноотстоящих точек и случай зависимости от многих переменных, хотя некоторые результаты для правильных сеток имеются, например в [5].

Сформулированный же нами подход является весьма общим применительно к регрессионным задачам и задачам аппроксимации, но не к задачам анализа случайных процессов и полей. Его пользу мы постараемся также проиллюстрировать рядом численных примеров.

3. Численный эксперимент. Далее приводятся численные результаты эксперимента для функции одной переменной, а в следующих разделах — для функции с двумя переменными. Эксперимент проводился в простейшем случае разностного оператора. Этого достаточно, чтобы судить о пользе предлагаемого подхода.

Как уже отмечалось, ряд результатов, полученных в работах [4, 5], позволяет выбирать базисные функции регрессии на правильной решетке и с обучающим разностным оператором. Приводимые ниже численные примеры свидетельствуют о возможностях некоторых обобщений на базе изложенной в предыдущих разделах теории.

3.1. Численный эксперимент с одной переменной. Отметим прежде всего возможность применения теоретических результатов в одномерном случае. Обсуждаются задача восстановления пропущенных наблюдений на равномерной сетке и аппроксимация в случае сетки с неравноотстоящими узлами с использованием операторного метода.

В качестве базисных функций f_j рассматриваются собственные функции линейных разностных операторов порядка $j = 1, \ldots, m$, с помощью которых аппроксимируются функции

$$g_1(x) = x^5 - 4.8\,x^4 + 8.24\,x^3 - 6.082\,x^2 + 1.7655\,x - 0.1235,$$

$$g_2(x) = 5\sin x^2.$$

3.1.1. Равномерная сетка. Рассмотрим задачу восстановления пропущенных значений функции, заданной на равномерной сетке. Результаты расчетов для функции g_1 (полином 5-го порядка, $x \in [0, 2]$) приведены в табл. 1. Шаг сетки $h = 0.01$, m — количество базисных функций, N_m — количество пропущенных точек в процентах от общего количества точек. Для оценки регрессионной модели рассчитываются средняя квадратическая ошибка (MSE) и средняя абсолютная ошибка (MAE).

Таблица 1. **Аппроксимация функции g_1 с пропущенными значениями**

m	N_m	MSE	MAE	m	N_m	MSE	MAE
4	90 %	0.0008	0.011	5	85 %	0.0008	0.010
4	57 %	0.0004	0.011	5	45 %	0.0001	0.008
4	0.5 %	0.0002	0.009	5	0.5 %	0.0001	0.009

Как показывают расчеты для функции g_1, даже при 85 % пропущенных значений аппроксимация операторным методом дает сравнительно невысокое значение MSE. Так как пропущенные значения формируются случайным образом, аппроксимирующие функции могут иметь различные аналитические выражения, но это существенно не влияет на оценки модели. Для генерации последовательности случайных чисел из заданного диапазона используется встроенная функция пакета R.

Вторая рассматриваемая функция g_2 является быстро осциллирующей; для повышения точности аппроксимации рассмотрим ее на интервале $[0, 4]$. Приведем результаты расчетов для функции g_2 в табл. 2 при различных значениях m и N_m. Погрешность аппроксимации уменьшается при увеличении числа базисных функций.

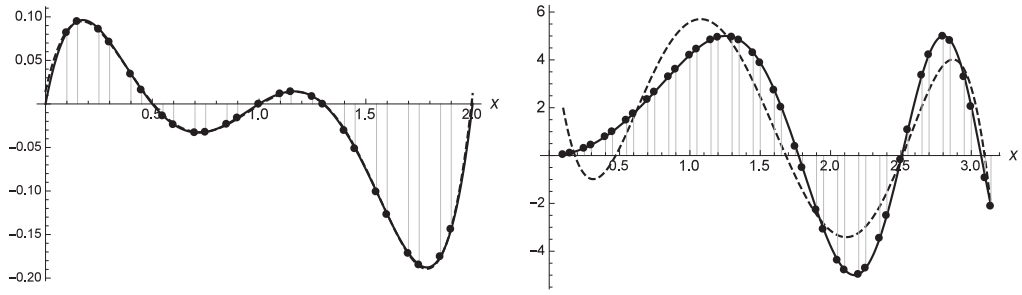
Таблица 2. **Аппроксимация функции g_2 с пропущенными значениями**

m	N_m	MSE	MAE	m	N_m	MSE	MAE
4	90 %	17.64	3.21	6	80 %	11.806	2.65
4	55 %	4.59	1.63	6	45 %	2.56	1.17
4	0.5 %	4.34	1.7	6	0.5 %	2.51	1.26

3.1.2. Неравномерная сетка. Аппроксимирующие функции находятся операторным методом. Узлы на заданном отрезке выбираются некоторым упорядоченным образом, например так: $[0.1, 0.15, 0.25, 0.3, 0.4, 0.45, 0.55, 0.6, 0.7, \ldots, 1.75, 1.85, 1.9]$. Приведем ниже базисные функции для g_1, g_2 соответственно:

$$\begin{pmatrix} 1.13^x \cos 0.42x \\ -1.13^x \sin 0.42x \\ 0.91^x \cos 0.68x \\ 0.91^x \sin 0.68x \\ 1.29^x \cos 0.42x \\ 1.29^x \sin 0.42x \end{pmatrix}, \quad \begin{pmatrix} 1.04^x \cos 0.33x \\ -1.04^x \sin 0.33x \\ 1.11^x \cos 0.59x \\ 1.11^x \sin 0.59x \\ 1.01^x \cos 0.32x \\ 1.01^x \sin 0.32x \end{pmatrix}.$$

На рис. 1 изображены тестовые g_1, g_2 (сплошная линия) и аппроксимирующие функции $\hat{g}_1, \hat{g}_2$ (пунктир), количество базисных функций m равно 6. Точками обозначены значения исходных функций в узлах.

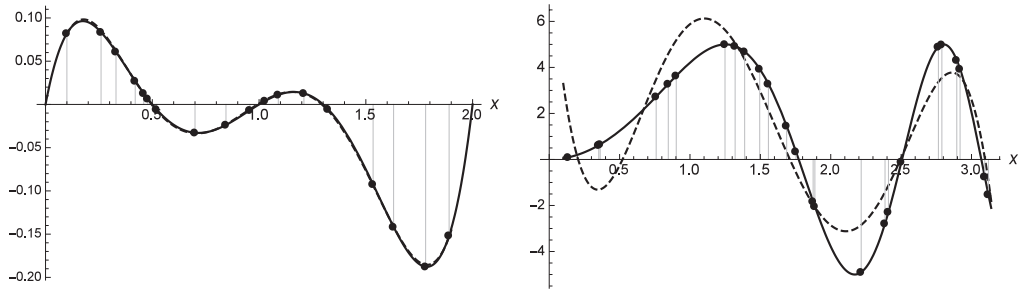


а *б*

Рис. 1. Аппроксимация функций g_1 (*а*) и g_2 (*б*) на неравномерной сетке.

Оценки аппроксимации для функций g_1, g_2 таковы: $\mathrm{MSE}_1 = 1.2 \cdot 10^{-6}$, $\mathrm{MSE}_2 = 1.2$. Графики функций $g_1, \hat{g}_1$ на рис. 1 практически совпадают.

3.1.3. Случайные узлы. В данном случае узлы на заданном отрезке выбираются случайным образом. На рис. 2 приведены результаты аппроксимации двух функций, рассмотренных выше. Количество базисных функций $m = 6$. Число точек для первого набора $N = 18$, для второго набора — $N = 25$. Оценки аппроксимации: $\mathrm{MSE}_1 = 1.3 \cdot 10^{-5}, \mathrm{MAE}_1 = 1.0 \cdot 10^{-3}$ и $\mathrm{MSE}_2 = 1.3, \mathrm{MAE}_2 = 0.99$ для функций g_1, g_2 соответственно.



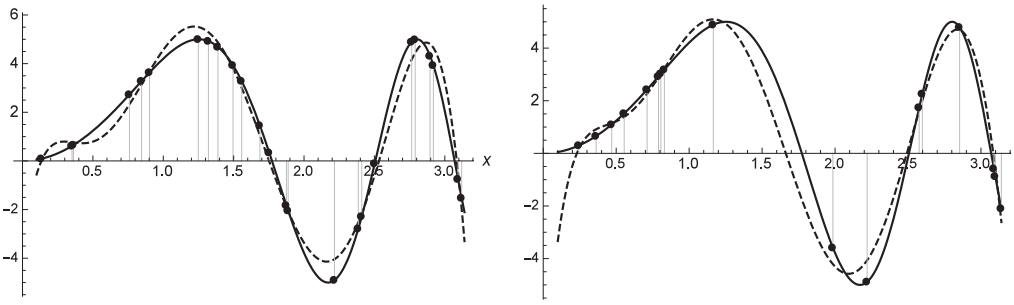
а *б*

Рис. 2. Аппроксимация функций g_1 (*а*) и g_2 (*б*) на случайных узлах.

Посмотрим, как выбор параметров влияет на качество аппроксимации. Полиномиальная функция хорошо аппроксимируется, поэтому проведем эксперимент для g_2. На рис. 3 приведены результаты расчетов при $m = 8$ для различных наборов случайных узлов. Оценки аппроксимации следующие: $\mathrm{MSE} = 0.15$, $\mathrm{MAE} = 0.32$ (рис. 3, *а*); $\mathrm{MSE} = 0.10$, $\mathrm{MAE} = 0.24$ (рис. 3, *б*). Для функции g_2 лучшей оказалась модель аппроксимации, построенная на случайных узлах.

Таким образом, используя операторный метод, можно управлять двумя параметрами в целях улучшения качества аппроксимации: количеством базисных функций и количеством узлов.

3.2. Аппроксимация функций двух переменных. Качество аппроксимации зависит от вида исходной функции. Рассмотрим несколько тестовых функций двух переменных — алгебраический полином, тригонометрическую функцию и су-



а *б*

Рис. 3. Аппроксимация функции g_2 на случайных узлах.

перпозицию тригонометрических функций и многочленов:

$$z_1(x,y) = -x^2 + 2y^2 - 4xy + 6x + 1, \quad x \in [0,3], \quad y \in [0,3],$$

$$z_2(x,y) = \sin(x+2y), \quad x \in [0,\pi], \quad y \in [0,\pi],$$

$$z_3(x,y) = -x \sin y \cos x, \quad x \in [0,6], \quad y \in [0,6],$$

$$z_4(x,y) = 2 \sin(x^2+y^2), \quad x \in [0,3], \quad y \in [0,3].$$

3.2.1. Аппроксимация на правильной сетке. Для задач аппроксимации на правильной сетке реализованы два метода: операторный (ОМ), где в качестве аппроксимирующих операторов выбраны линейные разностные операторы, и метод, использующий алгебраические полиномы в качестве базовых функций (ПМ).

Результаты расчетов для функции z_1 приведены в табл. 3, где h — шаг сетки, m — количество базисных функций. Уменьшение шага не оказывает существенного влияния на оценку точности операторного метода. Базисные функции для ОМ при $h = 0.01$ и $h = 0.006$ практически совпадают:

$$\{0.36^y \cos 0.71x,\ 2.75^y \cos 0.71x,\ 0.36^y \sin 0.71x,\ 2.75^y \sin 0.71x\}, \quad h = 0.01,$$

$$\{0.36^y \cos 0.70x,\ 2.72^y \cos 0.70x,\ 0.36^y \sin 0.70x,\ 2.72^y \sin 0.7x\}, \quad h = 0.006.$$

Таблица 3. **Аппроксимация функции** z_1

Метод	h	m	MSE	MAE	h	m	MSE	MAE
ОМ	0.01	4	0.71	0.58	0.006	4	0.7	0.57
ПМ	0.01	6	0	0	0.006	6	0	0

Таким образом, для достижения сравнимой точности методу ОМ потребуется значительно большее количество базисных функций.

Для тригонометрической функции z_2 базисные функции, найденные методом ОМ при $m = 4$, имеют вид $\{\sin x \sin 2y, \sin 2y \cos x, \cos x \cos 2y, \sin x \cos 2y\}$. Оценки аппроксимации для ОМ: MSE $= 0$, MAE $= 4.8 \cdot 10^{-9}$.

Базисные функции для ПМ — полиномы до шестой степени включительно, количество базисных функций $m = 20$. Оценки аппроксимации для ПМ: MSE $= 8.9 \cdot 10^{-5}$, MAE $= 6.9 \cdot 10^{-3}$.

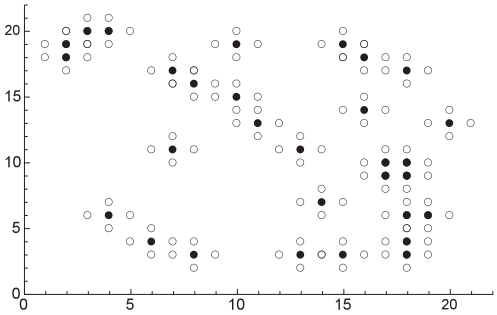


Рис. 4. Расчетные точки: • — случайные точки, ○ — пограничные точки.

Таким образом, тригонометрические функции типа z_2 хорошо аппроксимируются операторным методом.

Рассмотрим функции общего вида z_3 и z_4. Результаты расчетов приведены в табл. 4. Уменьшение шага сетки не оказывает существенного влияния на оценку точности операторного метода для функции z_3. Оценки ошибок для методов ОМ и ПМ практически совпадают. Однако при сравнимых оценках для ПМ потребуется 15 базовых функций, а для МО — только 4.

Таблица 4. **Аппроксимация функций** z_3, z_4

Метод	h	m	MSE	MAE	h	m	MSE	MAE
Аппроксимация функции z_3								
ОМ	0.01	4	0.87	0.68	0.006	4	0.87	0.68
ПМ	0.01	15	0.89	0.74	0.006	15	0.89	0.73
Аппроксимация функции z_4								
ОМ	0.01	4	1.94	1.25	0.05	4	1.94	1.25
ПМ	0.01	6	1.79	1.18	0.05	6	1.79	1.18

3.2.2. Аппроксимация с пропущенными значениями. На основе исходного набора данных формируется случайная выборка значений функции. Так как при численной реализации используется четырехточечный шаблон [6], к случайным точкам добавляются граничные точки. Все остальные точки считаются пропущенными. Пример формирования точек показан на рис. 4.

Пропущенные значения восстанавливаются по аппроксимирующей функции, которая является линейной комбинацией базисных функций.

Результаты расчетов для функций z_1 и для функций z_2, z_3, z_4 приведены в табл. 5 и 6 соответственно. Шаг сетки $h = 0.01$, количество базисных функций $m = 4$, N_m — количество пропущенных точек в процентах от N. Метрики MSE и MAE рассчитываются по восстановленной функции. Для сравнения приводятся расчеты при $N_m = 0\,\%$, то есть при отсутствии пропущенных значений.

Приведем для примера аналитический вид аппроксимирующей функции для z_1:

$$\sin 0.7x\,(8.84 \cdot 0.37^y - 0.12 \cdot 2.69^y) + \cos 0.7x\,(2.69^y - 0.03 \cdot 0.37^y).$$

Таблица 5. **Аппроксимация функции z_1 с пропущенными значениями**

Метод	h	N	N_m	MSE	MAE
ОМ	0.01	90 601	0 %	0.71	0.58
ОМ	0.01	90 601	44 %	0.70	0.57
ОМ	0.01	90 601	76 %	0.67	0.56
ОМ	0.01	90 601	95 %	0.68	0.55
ОМ	0.01	90 601	99 %	0.75	0.61

Таблица 6. **Аппроксимация функций z_2, z_3, z_4 с пропущенными значениями**

Метод	h	N	N_m	MSE	MAE
Аппроксимация функции z_2					
ОМ	0.01	90 601	0 %	0	$4.8 \cdot 10^{-9}$
ОМ	0.01	90 601	50 %	0	$5.5 \cdot 10^{-9}$
Аппроксимация функции z_3					
ОМ	0.01	361 201	0 %	0.87	0.68
ОМ	0.01	361 201	15 %	0.87	0.68
ОМ	0.01	361 201	30 %	0.87	0.67
Аппроксимация функции z_4					
ОМ	0.01	90 601	0 %	1.94	1.25
ОМ	0.01	90 601	25 %	1.94	1.25
ОМ	0.01	90 601	50 %	1.94	1.25
ОМ	0.01	90 601	75 %	1.95	1.25
ОМ	0.01	90 601	99 %	2.03	1.25

Таким образом, при аппроксимации методом ОМ значение ошибки определяется на полном наборе данных без пропущенных значений. При увеличении количества пропущенных значений до некоторого предела базисные функции отличаются незначительно.

3.2.3. Аппроксимация со случайной погрешностью. Рассмотрим влияние случайной ошибки на качество аппроксимации. Добавим к исходным данным аддитивный белый шум с дисперсией σ^2. Расчеты проводились при различных значениях шага сетки h и дисперсии σ^2. В табл. 7 приведены расчеты при различных уровнях защумленности, $\sigma = 0$ соответствует данным без случайной погрешности. Очевидно, что при увеличении σ^2 ошибка аппроксимации возрастает.

Таблица 7. **Аппроксимация функции z_1 со случайной ошибкой**

Метод	h	σ	MSE	MAE	h	σ	MSE	MAE
ОМ	0.01	0.001	34.07	4.89	0.03	0.001	5.13	1.54
ОМ	0.01	0.0001	3.76	1.34	0.03	0.0001	0.75	0.59
ОМ	0.01	0	0.71	0.58	0.03	0	0.74	0.59

Представляет интерес сравнение числа параметров аппроксимации для рассмотренных методов при одинаковом уровне ошибки для функции z_3. В табл. 8 приведены результаты расчетов, где m — число базисных функций, $h = 0.01$. Заметим, что число параметров, требуемых для метода ПМ, больше, чем для ОМ. И эта разница возрастает с уменьшением уровня шума.

Таблица 8. **Аппроксимация функции z_3 со случайной погрешностью**

Метод	σ	m	MSE	MAE	σ	m	MSE	MAE
ОМ	0.001	4	2.83	1.19	0.0001	4	2.80	1.20
ПМ	0.001	6	2.64	1.26	0.0001	6	2.63	1.26
ОМ	0.00001	4	0.88	0.69	0	4	0.87	0.68
ПМ	0.00001	14	0.92	0.74	0	15	0.89	0.73

Для быстроменяющейся функции z_4 качество аппроксимации методами ОМ и ПМ сравнимо при различных уровнях зашумленности (см. табл. 9).

Таблица 9. **Аппроксимация функции z_4 со случайной погрешностью**

Метод	σ	m	MSE	MAE	σ	m	MSE	MAE
ОМ	0.001	4	1.99	1.27	0.0001	4	1.94	1.25
ПМ	0.001	6	1.78	1.18	0.0001	6	1.79	1.18

4. Выводы. В данной работе продемонстрированы возможности применения операторного метода к задачам регрессионного анализа. В качестве аппроксимирующих операторов выбраны линейные разностные операторы с постоянными коэффициентами. Однако возможны обобщения: операторный метод позволяет применять линейные разностные операторы с переменными коэффициентами, линейные дифференциальные операторы, операторы осреднения.

Отметим ряд преимуществ данного метода. Предложенный метод дает возможность использовать меньшее число параметров (базисных функций) по сравнению, например, с полиномиальной регрессией. Кроме того, метод позволяет выбирать разные метрики, а не только метод наименьших квадратов.

При этом операторный метод имеет некоторые ограничения. Не исследованы возможности его применения для моделирования случайных процессов и построения прогнозных моделей. Сузив задачу, мы получили некоторые преимущества. В частности, в отличие от метода SSA-Гусеница, дающего близкие результаты на правильной сетке [7], операторный метод дает возможность строить регрессионные модели на неравномерных сетках для любого количества измерений.

Литература

1. Дрейпер Н., Смит Г. *Прикладной регрессионный анализ*, пер. с англ. 3-е изд. Киев, Диалектика (2016).

2. Гохберг И. Ц., Крейн М. Г. *Введение в теорию линейных несамосопряженных операторов*, пер. с англ. Москва, Наука (1965).

3. Донской В. И. Машинное обучение и обучаемость: сравнительный обзор. *Intellectual Archive*, № 933, 1–19 (2012).

4. Усевич К. Д. Разложение функций в двумерном варианте метода «Гусеница»-SSA и связанные с ним системы уравнений в частных производных. *Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления*, вып. 3, 151–160 (2009).

5. Ермаков С. М., Котова Л. Ю. О выборе базисных функций в регрессионном анализе. В: *Сб. работ кафедры статистического моделирования СПбГУ*, 3–43 (1999).

6. Самарский А. А. *Теория разностных схем.* Москва, Наука (1989).

7. Голяндина Н. Э., Усевич К. Д. Метод 2D-SSA для анализа двумерных полей. В: *Труды VII Международной конференции «Идентификация систем и задачи управления» SICPRO'08*, Москва, 1657–1727 (2008).

Статья поступила в редакцию 16 июля 2021 г.;
доработана 25 августа 2021 г.;
рекомендована к печати 2 сентября 2021 г.

Контактная информация:

Ермаков Сергей Михайлович — д-р физ.-мат. наук, проф.; sergej.ermakov@gmail.com
Леора Светлана Николаевна — канд. физ.-мат. наук, доц.; leora2008@mail.ru

On the choice of basic regression functions and machine learning

S. M. Ermakov[1], *S. N. Leora*[2]

[1] St Petersburg State University, 7–9, Universitetskaya nab., St Petersburg, 199034, Russian Federation
[2] St Petersburg State University of Economics,
30–32, nab. kanala Griboedova, St Petersburg, 191023, Russian Federation

For citation: Ermakov S. M., Leora S. N. On the choice of basic regression functions and machine learning. *Vestnik of Saint Petersburg University. Mathematics. Mechanics. Astronomy*, 2022, vol. 9 (67), issue 1, pp. 11–22. https://doi.org/10.21638/spbu01.2022.102 (In Russian)

As is known, the regression analysis task is widely used in machine learning problems, which allows to establish relationship between observed data and compactly store of information. Most often, a regression function is described by a linear combination of some of the selected functions $f_j(X)$, $j = 1, \ldots, m$, $X \in D \subset R^s$. If the observed data contains a random error, then the regression function restored from the observed data contains a random error and a systematic error depending on the selected functions f_j. The article indicates the possibility of optimal selection of functions f_j in the sense of a given functional metric, if it is known that the true dependence is consistent with some functional equation. In some cases (regular grids, $s \leq 2$), similar results can be obtained using the random process analysis method. The numerical examples given in this article illustrate much more opportunities for the task of constructing the regression function.

Keywords: regression analysis, approximation, basis functions, operator method, machine learning.

References

1. Draper N., Smith H. *Prikladnoi regressionnyi analiz.* 3rd ed. Kiev, Dialectica Publ. (2016). (In Russian) [Eng. transl.: Draper N., Smith H. *Applied Regression Analysis.* 3rd ed. New York, Wiley (1998)].

2. Gokhberg I. Ts., Kreyn M. G. *Vvedenie v teoriiu lineinykh nesamosopriazhennykh operatorov.* Moscow, Nauka Publ. (1965). (In Russian) [Eng. transl.: Gokhberg I. Ts., Kreyn M. G. *Introduction to the theory of linear non-self-adjoint operators in a Hilbert space.* In Ser.: Translations of Mathematical Monographs, vol. 18, AMS (1969)].

3. Donskoy V. I. Machine Learning and Learnability: Comparative Survey. *Intellectual Archive*, no. 933, 1–19 (2012). (In Russian)

4. Usevich K. D. Decomposition of functions in 2D-extension of SSA and related partial differential systems of equations. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, iss. 3, 151–160 (2009). (In Russian)

5. Ermakov S. M., Kotova L. Yu. On the choice of basic functions in regression analysis. In: *Collection of works of the Department of Statistical Modeling of St Petersburg State University*, 3–43 (1999). (In Russian)

6. Samarskiy A. A. *The theory of difference schemes*. Moscow, Nauka Publ. (1989). (In Russian)
7. Golyandina N. E., Usevich K. D. 2D-SSA Method for analysis of two-dimensional fields. In: *Proceedings of the VII International Conference "System Identification and Control Problems" SICPRO'08*, Moscow, 1657–1727 (2008). (In Russian)

Received: July 16, 2021
Revised: August 25, 2021
Accepted: September 2, 2021

Authors' information:

Sergey M. Ermakov — sergej.ermakov@gmail.com
Svetlana N. Leora — leora2008@mail.ru