ЧЕБЫШЕВСКИЙ СБОРНИК
Том 23. Выпуск 4.

УДК 519.85 DOI 10.22405/2226-8383-2022-23-4-52-63

Полный метод чебышевской интерполяции в задаче построения линейной регрессии

В. А. Горелик, Т. В. Золотова

Горелик Виктор Александрович — доктор физико-математических наук, ФИЦ ИУ РАН; профессор, Московский педагогический государственный университет (г. Москва).
e-mail: vgor16@mail.ru
Золотова Татьяна Валерьяновна — доктор физико-математических наук, профессор, Финансовый университет при Правительстве РФ (г. Москва).
e-mail: tgold11@mail.ru

Аннотация

Рассматривается линейная задача регрессионного анализа в предположении наличия шумов в выходной и входных переменных. Эта задача аппроксимации может интерпретироваться как несобственная задача интерполяции, для которой требуется оптимальным образом скорректировать положения исходных точек в пространстве данных так, чтобы они все лежали на одной гиперплоскости. Для оценки меры коррекции исходных данных используется минимаксный критерий, поэтому предлагаемый подход может быть назван полным методом чебышевской аппроксимации (интерполяции). Он приводит к нелинейной задаче математического программирования, которая сводится к решению конечного числа задач линейного программирования. Это число зависит экспоненциально от количества параметров, поэтому предлагаются некоторые методы преодоления данной проблемы. Полученные результаты иллюстрируются практическими примерами, основанными на реальных данных, а именно, проанализирован показатель рождаемости в Федеральных округах РФ в зависимости от таких факторов, как численность городского населения, доходы и инвестиции. Построены линейные регрессионные зависимости для двух и трех признаков. На эмпирическом факте статистической устойчивости (сохранение знаков коэффициентов) продемонстрирована возможность сокращения перебора задач линейного программирования.

Ключевые слова: обработка данных, линейная регрессия, матричная коррекция, чебышевская аппроксимация, задача линейного программирования.

Библиография: 15 названий.

Для цитирования:
В. А. Горелик, Т. В. Золотова. Полный метод чебышевской интерполяции в задаче построения линейной регрессии // Чебышевский сборник, 2022, т. 23, вып. 4, с. 52–63.

CHEBYSHEVSKII SBORNIK
Vol. 23. No. 4.

UDC 519.85 DOI 10.22405/2226-8383-2022-23-4-52-63

The total method of Chebyshev interpolation in the problem of constructing a linear regression

V. A. Gorelik, T. V. Zolotova

Gorelik Victor Alexandrovich — doctor of physical and mathematical sciences, FRC CSC RAS; professor, Moscow State Pedagogical University (Moscow).
e-mail: vgor16@mail.ru
Zolotova Tatiana Valerianovna — doctor of physical and mathematical sciences, professor, Financial University under the Government of the Russian Federation (Moscow).
e-mail: tgold11@mail.ru

Abstract

A linear problem of regression analysis is considered under the assumption of the presence of noise in the output and input variables. This approximation problem may be interpreted as an improper interpolation problem, for which it is required to correct optimally the positions of the original points in the data space so that they all lie on the same hyperplane. The minimax criterion is used to estimate the measure of correction of the initial data; therefore, the proposed approach can be called the total method of Chebyshev approximation (interpolation). It leads to a nonlinear mathematical programming problem, which is reduced to solving a finite number of linear programming problems. This number depends exponentially on the number of parameters, therefore, some methods are proposed to overcome this problem. The results obtained are illustrated with practical examples based on real data, namely, the birth rate in the Federal Districts of the Russian Federation is analyzed depending on factors such as urban population, income and investment. Linear regression dependencies for two and three features are constructed. Based on the empirical fact of statistical stability (conservation of signs of the coefficients), the possibility of reducing the enumeration of linear programming problems is demonstrated.

Keywords: data processing, linear regression, matrix correction, minimax criterion, linear programming problem.

Bibliography: 15 titles.

For citation:
V. A. Gorelik, T. V. Zolotova, 2022, “The total method of Chebyshev interpolation in the problem of constructing a linear regression” , *Chebyshevskii sbornik*, vol. 23, no. 4, pp. 52–63.

1. Введение

Методы коррекции несобственных и неустойчивых задач получили в настоящее время широкое распространение. Матричная коррекция (по совокупности всех исходных данных) применяется к несовместным системам линейных алгебраических уравнений и неравенств и несобственным задачам линейного программирования ([1-5]). Использование методов коррекции в задачах обработкой экспериментальных данных при наличии шумов во входных и выходных данных привели к появлению полного метода наименьших квадратов (TLS в англоязычной терминологии).

Задачу регрессионного анализа можно рассматривать как несобственную задачу интерполяции, которая заключается в построении функции $f : X \to Y$ из некоторого фиксированного класса функций Φ, такой, что поверхность, ею описываемая, точно проходит через имеющиеся точки исходных данных $(x^1, y^1), \ldots, (x^m, y^m)$, т. е.

$$y^i = f\left(x^i\right),\ i = 1, \ldots, m,\ f \in \Phi.$$

В связи с тем, что данные зачастую получены экспериментальным путем, задача становится несобственной (не имеющей решения). В этом случае рассматривается задача оптимальной коррекции (аппроксимации). Необходимо найти функцию, которая вместе с некоторым набором данных $[X_H, y_h]$ является решением задачи интерполяции, и этот набор данных является «ближайшим» к исходным точкам $[X, y]$ среди всех допустимых параметров, при которых задача интерполяции имеет решение. Эта задача аппроксимации формализуется введением некоторой матричной нормы и нахождением минимальной по данной норме матрицы коррекции $[X_H, y_h] - [X, y]$.

Если в качестве меры аппроксимации используется метрика l_2 (евклидова норма вектора и норма Фробениуса матрицы), то получаем полный метод наименьших квадратов. Он имеет вероятностное обоснование как метод максимального правдоподобия при использовании гипотезы нормального распределения ошибок как при измерении векторного аргумента x, так и значений функции y.

Нормальный закон распределения часто используется при моделировании случайных процессов. Это объясняется и удобством его применения при исследовании случайных процессов, и полезными свойствами нормального закона (например, устойчивостью). Видимо поэтому большинство современных работ по регрессионному анализу посвящено различным обобщениям и модификациям полного метода наименьших квадратов (см., например, [6-12]) и меньше внимания уделяется другим мерам аппроксимации.

Однако в ряде случаев, распределения случайных показателей отличаются от нормального. Отклонение гипотезы «нормальности» связано с тем, что значение коэффициента вытянутости (эксцесса) больше у статистических распределений, которые соответствуют реальным данным. Известно, что коэффициент вытянутости определяется через четвертый момент.

Это обстоятельство позволяет говорить о том, что такие распределения случайных величин имеют «тяжелые хвосты», т. е. соответствующая плотность распределения медленно убывает при $|x| \to \infty$ по сравнению с нормальной плотностью. Отклонение от нормального (гауссова) распределения случайных величин наблюдается в финансово-экономической области и характерно, например, для обменных курсов валют, для цен и доходностей акций. Это подтверждается как видом эмпирических плотностей (гистограмм), так и стандартными статистическими приемами обнаружения отклонений от нормального распределения [13].

Если распределение шумов в статистических данных отличается от нормального распределения, то метод наименьших квадратов теряет свое вероятностное обоснование. При этом в полной форме он достаточно сложен в вычислительном плане (приводит к нахождению собственных векторов или сингулярному разложению матриц).

В работе [5] было показано, что метод максимального правдоподобия при использовании гипотезы экспоненциального распределения шумов приводит к полиэдральной норме l_1, а метод построения линейной регрессии сводится при этом к решению совокупности задач линейного программирования. В работе [14] эти результаты были обобщены на задачу совместного преобразования данных и аппроксимации, что приводит к новому классу задач параметрической коррекции.

В данной работе предлагается использование минимаксного критерия аппроксимации, т.е. использование нормы l_∞ матрицы коррекции. Так как минимаксный критерий принято связывать с именем П.Л. Чебышева, то предлагаемый подход естественно назвать полным методом чебышевской аппроксимации.

2. Задача коррекции несовместной системы линейных уравнений в метрике l_∞

Рассмотрим в качестве вспомогательной задачу коррекции несовместной системы линейных уравнений $Bz = c$, где вектор z имеет размерность n, вектор c – размерность m, матрица B имеет размерность $m \times n$. Несовместная система обычно является переопределенной, т.е. $m > n$. Обозначим компоненты вектора c через c_i, строки матрицы B – через b^i, $i = 1, \ldots, m$.

Введем матрицу H и вектор h соответствующих размерностей так, что система $(B+H)z = c + h$ становится совместной, т. е. множество ее решений Z непусто. Поставим задачу минимальной коррекции как задачу минимизации некоторой нормы расширенной матрицы $\overline{H} = [-h \;\; H]$ при условии $Z \neq \emptyset$:

$$\min_{z,H,h} \{ \| [-h \;\; H] \| \mid (B + H)\, z = c + h \}. \tag{1}$$

Матрица $\overline{H}$ имеет размерность $m \times (n+1)$. Обозначим ее элементы через h_{ij}, $i = 1, \ldots, m$, $j = 0, 1, \ldots, n$, строки матрицы $\overline{H}$ – через h^i.

Норма l_∞ матрицы по определению есть

$$\|A\|_{l_\infty} = \max_{i,j} |a_{ij}| \; .$$

Для того, чтобы получить решение задачи (1) в данной норме, найдем выражение этой нормы через векторные нормы. Для этого нам будет полезно следующее определение.

Обобщенная φ, ψ – норма произвольной матрицы A есть

$$\|A\|_{\varphi,\psi} = \max_{z \neq 0} \frac{\psi\,(Az)}{\varphi(z)},$$

где φ, ψ – некоторые векторные нормы. Покажем, что норма матрицы A в метрике l_∞ является частным случаем обобщенной матричной нормы, а именно,

$$\|A\|_{1,\infty} = \|A\|_{l_\infty} = \max_{i,j} |a_{ij}| \; .$$

Действительно, по определению

$$\|A\|_{1,\infty} = \max_{z \neq 0} \frac{\|Az\|_\infty}{\|z\|_1} \; = \max_{z \neq 0} \frac{\max\limits_i \left| \sum_j a_{ij} z_j \right|}{\sum_j |z_j|}$$

(для краткости здесь соответствующие нормы векторов обозначены 1 и ∞). Очевидно, данную задачу на максимум отношения можно заменить на условный экстремум:

$$\begin{gathered} \max_i \left| \sum_j a_{ij} z_j \right| \; \to \max_z \; , \\ \sum_j |z_j| = 1. \end{gathered}$$

Максимум линейной функции достигается в вершинах допустимого множества, поэтому при фиксированном i максимум по z выражения $\sum_j a_{ij} z_j$ равен $\max\limits_j |a_{ij}| \; = |a_{ij_0}|$ и достигается на векторе z, у которого $z_{j_0} = sign\,(a_{ij_0})$, а все остальные компоненты равны нулю. Впрочем, речь идет о максимизации модуля суммы, поэтому можно положить $z_{j_0} = 1$. Далее, так как операции взятия максимума переставимы, то равенство $\|A\|_{1,\infty} = \|A\|_{l_\infty}$ доказано.

В работе [4] доказано, что минимум нормы $\|A\|_{\varphi,\psi}$ у матрицы A, являющейся при фиксированном z решением системы уравнений $Az = b$, равен

$$\|A\|_{\varphi,\psi} = \frac{\psi(b)}{\varphi(z)}.$$

Применим этот результат для уравнения $(B+H)z = c+h$ задачи (1). Преобразуем уравнение к виду

$$[-h \quad H]\,(1,\ z) = c - Bz.$$

Тогда минимальное значение обобщенной нормы расширенной матрицы коррекции, для которой некоторое фиксированное z удовлетворяет условию в задаче (1), есть

$$\|[-h \quad H]\|_{\varphi,\psi} = \frac{\psi(c-Bz)}{\varphi(1,z)}.$$

Таким образом, при фиксированном z минимальная в этой метрике норма расширенной матрицы равна

$$\max_{ij}|h_{ij}| \ = \frac{\max\limits_{1\leqslant i\leqslant m}\left|c_i - b^i z\right|}{1+\sum_{j=1}^{n}|z_j|}. \tag{2}$$

Задача минимальной коррекции (1) при этом сводится к минимизации отношения в правой части (2) по переменной z:

$$h^0 = \min_z \frac{\max\limits_{1\leqslant i\leqslant m}\left|c_i - b^i z\right|}{1+\sum_{j=1}^{n}|z_j|}.$$

Далее применим этот результат к решению задачи построения линейной регрессии.

3. Полная задача построения линейной регрессии в метрике l_∞

Математическая постановка задачи построения линейной регрессии заключается в следующем. Исходные данные, описывающие зависимость величины y от вектора переменных x, представляют собой множество точек $(x_1^1,\ldots,x_n^1,y^1),\ldots,(x_1^m,\ldots,x_n^m,y^m)$. Эти данные представим в виде информационной матрицы

$$[-y \quad X] = \begin{pmatrix} -y^1 & x_1^1 & \ldots & x_n^1 \\ -y^2 & x_1^2 & \ldots & x_n^2 \\ \ldots & \ldots & \ldots & \ldots \\ -y^m & x_1^m & \ldots & x_n^m \end{pmatrix}.$$

Рассматривается задача построения по заданным m точкам такой аффинной функции от n переменных $f: R^n \to R$ вида

$$f(x) = a_1x_1 + a_2x_2 + \cdots + a_nx_n + a_0 = \langle a, x\rangle + a_0, \tag{3}$$

что максимальный из модулей отклонений по всем координатам всех точек от определяемой ею гиперплоскости минимален. Сформулируем соответствующую задачу коррекции системы линейных уравнений. Условие принадлежности точек $(x^1,y^1),\ldots,(x^m,y^m)$ некоторой гиперплоскости L можно записать как

$$\langle a, x^i\rangle + a_0 = y^i, \quad i = 1,\ldots,m,$$

или в матричной форме

$$[X \quad e] \cdot \overline{a} = y, \tag{4}$$

где $y = \left(y^1,\ y^2,\ \ldots, y^m\right)^T$, $e = (1,\ 1,\ \ldots,\ 1)^T \in R^m$, $\overline{a} = (a, a_0)^T \in R^{n+1}$, X – матрица размера $m \times n$, строками которой являются векторы x^i.

Если через заданные точки нельзя провести гиперплоскость, то полученная система линейных уравнений (4) несовместна. Задача минимальной коррекции данной системы в норме l_∞ будет иметь следующий вид:

$$h^0 = \inf_{H,h,\overline{a}} \left\{ \|[-h \quad H]\|_{l_\infty} \mid [X + H \quad e]\,\overline{a} = y + h \right\}. \tag{5}$$

Задача (5) представляет собой задачу коррекции несовместной системы линейных уравнений. Правда в отличие от рассмотренного в предыдущем разделе случая здесь последний столбец, соответствующий фиктивной переменной, является фиксированным, однако это легко учитывается при применении формулы (2). Сформулируем теорему для критерия минимума полиэдральной нормы l_∞ матрицы коррекции, позволяющую получить решение задачи построения регрессии такого типа, т.е. нахождения оптимального значения коэффициентов $\overline{a}^0$.

Теорема 1. *Пусть в пространстве признаков R^n даны m точек $\left(x_1^1,\ x_2^1,\ \ldots,\ x_n^1\right)$, $\ldots$, $(x_1^m,\ x_2^m,\ \ldots,\ x_n^m)$, а в пространстве R множество ответов $y^1,\ \ldots, y^m$, и не существует аффинной функции (3) такой, что $y^i = f\left(x^i\right)$, $i = 1, \ldots, m$. Тогда задача нахождения минимального изменения информационной матрицы параметров $[-y \quad X]$ в смысле минимума нормы l_∞, в результате которого интерполяционная аффинная функция существует, эквивалентна задаче математического программирования*

$$\begin{gathered} u \to \min_{u,v,q,q_0}, \\ u \geqslant vy^i - x^i q - q_0,\ i = 1, \ldots, m, \\ u \geqslant -vy^i + x^i q + q_0,\ i = 1, \ldots, m, \\ v + \textstyle\sum_{j=1}^n |q_j| = 1, \\ u \geqslant 0,\ v \geqslant 0. \end{gathered} \tag{6}$$

Если существует решение задачи (6) $\left(u^0, v^0, q^0, q_0^0\right)$ такое, что $v^0 \neq 0$, то

$$h^0 = u^0,\ a^0 = \frac{q^0}{v^0},\ a_0^0 = \frac{q_0^0}{v^0}\ . \tag{7}$$

Доказательство. Используем формулу (2) для несовместной системы (4):

$$h^0 = \min_{a,a_0} \frac{\max\limits_{1 \leqslant i \leqslant m} \left|y^i - x^i a - a_0\right|}{1 + \sum_{j=1}^n |a_j|}.$$

Введем в рассмотрение скалярные переменные $v = \frac{1}{1+\sum_{j=1}^n |a_j|}$, $q_0 = va_0$, вектор $q = va$ и скалярную переменную u, удовлетворяющую условиям

$$u \geqslant vy^i - x^i q - q_0, \quad u \geqslant -vy^i + x^i q + q_0, \quad i = 1, \ldots, m.$$

При этом выполняются условия $u \geqslant 0$ и $v \geqslant 0$, но переменные q и q_0 могут иметь, вообще говоря, любой знак. В соответствии с формулой (2) необходимо минимизировать в новых переменных величину

$$\max_{1 \leqslant i \leqslant m} \left|vy^i - x^i q - q_0\right|\ ,$$

что эквивалентно минимизации переменной u при заданных на нее ограничениях. Таким образом, получаем задачу математического программирования (6), а с учетом введенных замен переменных и формулы (7). Теорема доказана. □

Задача (6) не является задачей линейного программирования. Сложность решения полученной задачи математического программирования связана с наличием модулей некоторых переменных. В случае рассмотрения нормы l_1 эта сложность преодолевалась сравнительно легко и привела, в конечном счете, к решению $2n$ задач линейного программирования [5]. Для задачи (6) дело обстоит несколько сложнее. Поэтому рассмотрим некоторые методы решения этой проблемы.

Задачу (6) также можно свести к решению конечного числа задач линейного программирования. Это можно сделать следующим способом. Предположим, что мы знаем априорно знаки коэффициентов регрессии, т.е. знаки компонент вектора $\overline{a}$. Тогда мы можем ввести для них новые переменные, которые совпадают с исходными переменными для положительных компонент и противоположного знака для отрицательных компонент. Это эквивалентно тому, что соответствующие столбцы матрицы X меняют знак, а все коэффициенты регрессии считаются неотрицательными. Тогда в задаче (6) элементы данных столбцов информационной матрицы меняют знаки, а модули просто исчезают.

Если из каких-то априорных соображений мы можем судить о знаках коэффициентов линейной регрессии, то задача (6) просто сводится к решению одной задачи линейного программирования. Однако в общем случае мы должны перебрать все возможные варианты знаков коэффициентов и из всех получающихся задач линейного программирования выбрать ту, которая дает наименьшее значение невязки u^0.

Таким образом, в общем случае задача (6) сводится к решению 2^n задач линейного программирования, т. е. рост экспоненциальный. Что можно сказать по этому поводу? Вообще говоря, число параметров (признаков) обычно на много меньше числа данных ($n \ll m$). При 10-20 параметрах проблемы перебора положительных и отрицательных значений коэффициентов нет. Кроме того, из содержательного смысла параметров зачастую можно судить хотя бы о части знаков коэффициентов. Если некоторые из них очевидны, то перебор уменьшается. Более подробно мы обсудим этот вопрос ниже при рассмотрении практических примеров. В общем же случае можно использовать метод ветвей и границ.

4. Построение демографического тренда в Российской Федерации

Рассмотрим практические пример, основанные на реальных данных, взятых с сайта Росстат [15]. Мы будем анализировать показатель рождаемости в Федеральных округах в зависимости от таких факторов, как численность городского населения, доходы и инвестиции. В таблице 1 приведены данные за 2019 год по Федеральным округам РФ: коэффициент рождаемости, удельный вес городского населения, доходы населения, инвестиции в области здравоохранения и социальных услуг.

Введем переменные x_1, x_2, x_3 – численность городского населения, доходы и инвестиции соответственно. По исходным данным, представленным в таблице 1, построим регрессионную зависимость в метрике l_∞ сначала от двух переменных: численности городского населения x_1 и доходов x_2.

Таблица 1: Анализируемые социально-экономические показатели по регионам РФ за 2019 г.

Федеральные округа	Общий коэффициент рождаемости (число родившихся на 1000 чел. населения)	Удельный вес городского населения (в %)	Среднедушевые доходы (в месяц, тыс. руб.)	Инвестиции в области здравоохранения и социальных услуг (млрд. руб.)
Центральный	9.3	82.3	46.9	91.3
Северо-Западный	9.6	84.9	37.9	26.6
Южный	9.8	62.8	29.9	31.6
Северо-Кавказский	13.7	50.3	24.4	14.7
Приволжский	9.6	72.2	28.3	41.9
Уральский	10.9	81.6	36.9	28.8
Сибирский	10.4	74.3	27.2	33.6
Дальневосточный	11.1	72.9	37.9	24.3

Таблица 2: Результаты вычислений при нахождении решения задачи (6) для двух переменных

u^0	v^0	q^0		q_0^0
		Знаки компонент вектора q^0	Значения компонент вектора q^0	
2.3	1	$++$	$(0,\ 0)$	12.7
1.443	1.009	$-+$	$(-0.103,\ 0.095)$	16.554
2	1.049	$+-$	$(0,\ -0.049)$	14.913
1.584	1.082	$--$	$(-0.082,\ 0)$	18.742

Информационная матрица имеет вид

$$[-y \quad X] = \begin{pmatrix} -y^1 & x_1^1 & x_2^1 \\ -y^2 & x_1^2 & x_2^2 \\ -y^3 & x_1^3 & x_2^3 \\ -y^4 & x_1^4 & x_2^4 \\ -y^5 & x_1^5 & x_2^5 \\ -y^6 & x_1^6 & x_2^6 \\ -y^7 & x_1^7 & x_2^7 \\ -y^8 & x_1^8 & x_2^8 \end{pmatrix} = \begin{pmatrix} -9.3 & 82.3 & 46.9 \\ -9.6 & 84.9 & 37.9 \\ -9.8 & 62.8 & 29.9 \\ -13.7 & 50.3 & 24.4 \\ -9.6 & 72.2 & 28.3 \\ -10.9 & 81.6 & 36.9 \\ -10.4 & 74.3 & 27.2 \\ -11.1 & 72.9 & 37.9 \end{pmatrix}.$$

Для этой информационной матрицы задача (6) имеет вид:

$$\begin{gathered} u \to \min_{u,v,q_0,q_1,q_2}, \\ u \geqslant vy^i - \left(x_1^i q_1 + x_2^i q_2\right) - q_0,\ i = 1, \ldots, 8, \\ u \geqslant -vy^i + \left(x_1^i q_1 + x_2^i q_2\right) + q_0,\ i = 1, \ldots, 8, \\ v + |q_1| + |q_2| = 1, \\ u \geqslant 0,\ v \geqslant 0. \end{gathered}$$

Решение данной задачи с модулями вектора переменных q предполагает решение $2^2 = 4$ задач линейного программирования, которые получаются путем перебора знаков компонент вектора переменных q. Результаты решения 4 задач линейного программирования представлены в таблице 2.

Определяем минимальное из четырех значение целевой функции и соответствующий набор оптимальных значений переменных:

$$u^0 = 1.443, v^0 = 0.009,\ q^0 = (-0.103,\ 0.077)\,, q_0^0 = 16.554.$$

Вычисляя по формулам (7) коэффициенты регрессии

$$a^0 = \frac{q^0}{v^0} = \frac{(-0.103,\ 0.077)}{0.009} = (-0.103,\ 0.094),$$

$$a_0^0 = \frac{q_0^0}{v^0} = \frac{16.554}{0.009} = 16.415,$$

получаем уравнение регрессии

$$y = -0.103x_1 + 0.094x_2 + 16.415.$$

Построим регрессионную зависимость в метрике l_∞ от трех переменных, значения которых представлены в таблице 1.

Информационная матрица примет вид

$$[-y\ \ X] = \begin{pmatrix} -y^1 & x_1^1 & x_2^1 & x_3^1 \\ -y^2 & x_1^2 & x_2^2 & x_3^2 \\ -y^3 & x_1^3 & x_2^3 & x_3^3 \\ -y^4 & x_1^4 & x_2^4 & x_3^4 \\ -y^5 & x_1^5 & x_2^5 & x_3^5 \\ -y^6 & x_1^6 & x_2^6 & x_3^6 \\ -y^7 & x_1^7 & x_2^7 & x_3^7 \\ -y^8 & x_1^8 & x_2^8 & x_3^8 \end{pmatrix} = \begin{pmatrix} -9.3 & 82.3 & 46.9 & 91.3 \\ -9.6 & 84.9 & 37.9 & 26.6 \\ -9.8 & 62.8 & 29.9 & 31.6 \\ -13.7 & 50.3 & 24.4 & 14.7 \\ -9.6 & 72.2 & 28.3 & 41.9 \\ -10.9 & 81.6 & 36.9 & 28.8 \\ -10.4 & 74.3 & 27.2 & 33.6 \\ -11.1 & 72.9 & 37.9 & 24.3 \end{pmatrix}.$$

Для этой информационной матрицы получаем задачу (6):

$$\begin{gathered} u \to \min_{u,v,q_0,q_1,q_2,q_3}, \\ u \geqslant vy^i - \left(x_1^i q_1 + x_2^i q_2 + x_3^i q_3\right) - q_0,\ i = 1, \ldots, 8, \\ u \geqslant -vy^i + \left(x_1^i q_1 + x_2^i q_2 + x_3^i q_3\right) + q_0,\ i = 1, \ldots, 8, \\ v + |q_1| + |q_2| + |q_3| = 1, \\ u \geqslant 0,\ v \geqslant 0. \end{gathered}$$

Решение данной задачи с модулями вектора переменных q предполагает решение $2^3 = 8$ задач линейного программирования, которые получаются путем перебора знаков компонент вектора переменных q. Результаты решения 8 задач линейного программирования представлены в таблице 3.

Определяем минимальное из восьми значение целевой функции и соответствующий набор оптимальных значений переменных:

$$u^0 = 1.207, v^0 = 0.97,\ q^0 = (-0.099,\ 0.202,\ -0.074)\,, q_0^0 = 13.202.$$

Вычисляя по формулам (7) коэффициенты регрессии

$$a^0 = \frac{q^0}{v^0} = \frac{(-0.099,\ 0.202,\ -0.074)}{0.97} = (-0.102,\ 0.209,\ -0.076),$$

$$a_0^0 = \frac{q_0^0}{v^0} = \frac{13.202}{0.97} = 13.607,$$

Таблица 3: Результаты вычислений при нахождении решения задачи (6) для трех переменных

u^0	v^0	q^0		q_0^0
		Знаки компонент вектора q^0	Значения компонент вектора q^0	
2.2	1	$+\,+\,+$	(0, 0, 0)	11.5
1.437	1.008	$-\,+\,+$	(−0.145, 0.136, 0)	16.328
2.052	1.016	$+\,-\,+$	(0, −0.016, 0)	12.273
1.812	1.061	$+\,+\,-$	(0, 0, −1.061)	13.618
1.601	1.103	$+\,-\,-$	(0, −0.057, −0.047)	15.589
1.207	0.97	$-\,+\,-$	(−0.099, 0.202, −0.074)	13.202
1.528	1.198	$-\,-\,+$	(−0.098, 0, 0)	18.460
1.395	1.115	$-\,-\,-$	(−0.087, 0, −0.028)	18.676

получаем уравнение регрессии

$$y = -0.102x_1 + 0.209x_2 - 0.076x_3 + 13.607.$$

При сравнении результатов построения уравнения регрессии для двух и трех переменных (факторов) обращает на себя внимание тот факт, что коэффициенты при соответствующих переменных близки по абсолютным значениям, а главное – совпадение их знаков. Этот факт не является удивительным, т. к. положительное или отрицательное влияние каждого фактора должно быть устойчивым, хотя это утверждение не является строгим. Данное эмпирическое утверждение можно использовать для сокращения вычислительных процедур. Если последовательно увеличивать количество параметров, то в предположении постоянства знака коэффициента для каждого фактора на следующем шаге достаточно решать две задачи (положительное и отрицательное значение коэффициента при добавляемом факторе). При такой последовательной процедуре построения регрессии от n переменных достаточно решать $2n$ задач линейного программирования вместо 2^n. В примере построения уравнения регрессии с тремя переменными достаточно решить задачи со знаками коэффициентов $(-++)$ и $(-+-)$, а всего с учетом примера с двумя переменными – 6 задач.

5. Заключение

В данной работе была рассмотрен подход к построению линейной регрессии как несобственной задачи интерполяции, основанный на матричной коррекции системы линейных уравнений, выражающей условие принадлежности всех точек пространства исходных данных одной гиперплоскости. В качестве меры коррекции (аппроксимации) использована норма матрицы l_∞. В геометрической интерпретации это означает минимизацию максимума модулей отклонений от гиперплоскости всех точек по всем координатам. Данный подход в вычислительном плане приводит к решению совокупности задач линейного программирования, однако в общем случае их число растет экспоненциально с ростом числа параметров. Предложены некоторые способы преодоления этой сложности. Приведен пример построения демографического тренда по реальным данным.

СПИСОК ЦИТИРОВАННОЙ ЛИТЕРАТУРЫ

1. Eremin I. I. Theory of linear optimization. Inverse and ill-posed problems series. VSP: Utrecht, Boston, Koln, Tokyo, 2002.

2. Горелик В. А. Матричная коррекция задачи линейного программирования с несовместной системой ограничений // Журн. вычисл. матем. и матем. физика. 2001. Т. 41, № 11. С. 1697-1705.

3. Горелик В. А., Ерохин В. И. Оптимальная матричная коррекция несовместных систем линейных алгебраических уравнений по минимуму евклидовой нормы. М.: ВЦ РАН, 2004.

4. Горелик В. А., Ерохин В. И. Печенкин Р. В. Численные методы коррекции несобственных задач линейного программирования и структурных систем уравнений. М.: ВЦ РАН, 2006.

5. Горелик В. А., Трембачева О. С. Решение задачи линейной регрессии с использованием методов матричной коррекции в метрике l_1 // Журн. вычисл. матем. и матем. физика. 2016, т. 56, №2. С. 202-207.

6. Back A. The matrix-restricted total least squares problem // Signal Process. 2007. Vol. 87, №10. P. 2303-2312.

7. Hnětynková I., Plešinger M., Sima D. M., Starakoš Z., Van Huffel S. The total least squares problem in $AX \approx B$: A new classification with the relationship to the classical works // SIAM Journal on Matrix Analysis and Applications. 2011. Vol. 32, issue 3. P. 748-770.

8. Hnětynková I., Plešinger M., Žáková J. On TLS formulation and core reduction for data fitting with generalized models // Linear Algebra and Its Applications. 2019. Vol. 577. P. 1-20.

9. Hnětynková I., Plešinger M., Žáková J. Solvability classes for core problems in matrix total least squares minimization // Applications of Mathematics. 2019. Vol. 64, issue 2. P. 103-128.

10. Markovsky I., Van Huffel S. Overview of total least-squares methods // Signal Processing. 2007. Vol. 87, issue 10. P. 2283-2302.

11. Meng L., Zheng B., Wei Y.Condition numbers of the multidimensional total least squares problems having more than one solution // Numerical Algorithms. 2020. Vol. 84, issue 3. P. 887-908.

12. Shklyar S. Consistency of the total least squares estimator in the linear errors-in-variables regression // Modern Stochastics: Theory and Applications. 2018. Vol. 5, issue 3. P. 247-295.

13. Ширяев А. Н. Основы стохастической финансовой математики. Т. 1. Факты. Модели. М.: МЦНМО, 2016.

14. Gorelik V. A., Zolotova T. V. Method of parametric correction in data transformation and approximation problems // Lecture Notes in Computer Science (LNCS). 2020. Vol. 12422. P.122-133.

15. Регионы России. Социально-экономические показатели 2019, https://rosstat.gov.ru. Дата обращения: 30 октября 2021.

REFERENCES

1. Eremin, I.I. 2002, *Theory of linear optimization. Inverse and ill-posed problems series*, VSP, Utrecht, Boston, Koln, Tokyo.

2. Gorelik, V. A. 2001, "Matrix correction of a linear programming problem with inconsistent constraints", *Computational mathematics and mathematical physics*, vol. 11, no. 41, pp. 1615-1622.

3. Gorelik, V. A., Erohin, V. I. 2004, *Optimal matrix correction of inconsistent systems of linear algebraic equations by minimal Euclidean norm*, CC RAS, Moscow.

4. Gorelik, V. A., Erokhin, V. I., Pechenkin, R. V. 2006, *Numerical methods for correcting improper linear programming problems and structural systems of equations*, CC RAS, Moscow.

5. Gorelik, V. A., Trembacheva, O. S. 2016, "Solution of the Linear Regression Problem Using Matrix Correction Methods in the l_1 Metric", *Computational Mathematics and Mathematical Physics*, vol. 56, no. 2, pp. 200-205.

6. Back A. 2007, "The matrix-restricted total least squares problem", *Signal Process*, vol. 87, no. 10, pp. 2303-2312.

7. Hnětynková I., Plešinger M., Sima D. M., Starakoš Z., Van Huffel S. 2011, "The total least squares problem in $AX \approx B$: A new classification with the relationship to the classical works", *SIAM Journal on Matrix Analysis and Applications*, vol. 32, issue 3, pp. 748-770.

8. Hnětynková I., Plešinger M., Žáková J. 2019. "On TLS formulation and core reduction for data fitting with generalized models", *Linear Algebra and Its Applications*, vol. 577, pp. 1-20.

9. Hnětynková I., Plešinger M., Žáková J. 2019. "Solvability classes for core problems in matrix total least squares minimization", *Applications of Mathematics*, vol. 64, issue 2, pp. 103-128.

10. Markovsky I., Van Huffel S. 2007 "Overview of total least-squares methods", *Signal Processing*, vol. 87, issue 10, pp. 2283-2302.

11. Meng L., Zheng B., Wei Y. 2020. "Condition numbers of the multidimensional total least squares problems having more than one solution", *Numerical Algorithms*, vol. 84, issue 3, pp. 887-908.

12. Shklyar S. 2018. "Consistency of the total least squares estimator in the linear errors-in-variables regression", *Modern Stochastics: Theory and Applications*, vol. 5, issue 3, pp. 247-295.

13. Shiryaev A. N. 2016, *Fundamentals of stochastic financial mathematics*, vol. 1, Facts, Models, MTSNMO, Moscow.

14. Gorelik V. A., Zolotova T. V. 2020, "Method of parametric correction in data transformation and approximation problems", *Lecture Notes in Computer Science*, vol. 12422, pp.122-133.

15. Regions of Russia. Socio-economic indicators 2019, https://rosstat.gov.ru (accessed 30 October 2021).

Получено: 13.07.2022
Принято в печать: 8.12.2022