

ЗАДАЧА ПОСТРОЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ С ПРЕОБРАЗОВАНИЕМ ДАННЫХ И ИСПОЛЬЗОВАНИЕМ МЕТРИКИ L_1

Горелик В. А.¹

*(Вычислительный центр им. А.А. Дородницына
ФИЦ ИУ РАН,*

Институт математики и информатики МПГУ, Москва)

Золотова Т. В.²

*(Финансовый университет
при Правительстве РФ, Москва)*

Рассматривается линейная задача регрессионного анализа в предположении наличия шумов в выходной и входных переменных, т.е. задача построения аффинной функции, которая в некотором смысле описывает наилучшим образом исходные данные. Геометрическая интерпретация заключается в построении гиперплоскости в $(n+1)$ -мерном пространстве, которая проходит ближе всех к заданным точкам в смысле выбранной меры близости. Эта задача аппроксимации может также интерпретироваться как несобственная (не имеющая решения) задача интерполяции, для которой требуется оптимальным образом изменить (скорректировать) положения точек так, чтобы они все лежали на одной гиперплоскости. Для оценки меры аппроксимации (коррекции) исходных данных использована метрика l_1 , которая имеет вероятностное обоснование как метод максимального правдоподобия при двустороннем экспоненциальном распределении шумов. Кроме того, рассматривается случай необходимости предварительного линейного преобразования данных, притом, что коррекции подлежит исходная информационная матрица. Поэтому возникающие задачи относятся к новому классу параметрической коррекции. Показано, что эти задачи аппроксимации (коррекции) сводится к совокупности конечного числа задач линейного программирования. Для работы с матрицами и решения вспомогательных оптимизационных задач использован математический пакет MatLab.

Ключевые слова: обработка данных, регрессионный анализ, матрица коррекция, метод максимального правдоподобия, экспоненциальное распределение.

¹ Виктор Александрович Горелик, д.ф.-м.н., профессор (vgor16@mail.ru).

² Татьяна Валерьяновна Золотова, д.ф.-м.н., профессор (tgold11@mail.ru).

1. Введение

Процесс математического моделирования и управления сложными системами в качестве необходимой составляющей включает процедуры предварительной обработки данных и методы регрессионного анализа.

Задача регрессионного анализа состоит в выборе функции, которая в некотором смысле достаточно точно описывает заданные таблично экспериментальные данные. Так как точное описание, как правило, невозможно, то ее можно интерпретировать как несобственную (не имеющую решения) задачу интерполяции. Последняя состоит в построении функции $f: X \rightarrow Y$ из некоторого класса Φ такой, что исходные данные $(x^1, y^1), \dots, (x^m, y^m)$ удовлетворяют системе уравнений

$$(1) \quad y^i = f(x^i), i = 1, \dots, m, f \in \Phi.$$

Геометрически это означает, что поверхность, описываемая данной функцией, проходит через заданные точки. Так как данные обычно получены экспериментальным путем и включают неточности измерения или шумы, то эта задача, как правило, не имеет решения для фиксированного класса функций. В таком случае рассматривается задача оптимальной коррекции (аппроксимации): требуется найти функцию, которая удовлетворяет условию (1) для некоторых других данных $[X_H, y_h]$, причем эти данные являются в некотором смысле ближайшими к исходным $[X, y]$. Эта задача формализуется путем введения меры близости как некоторой матричной нормы и нахождения минимальной в смысле этой нормы матрицы коррекции $A = [X_H, y_h] - [X, y]$.

Методы матричной коррекции (по совокупности всех исходных данных) получили в настоящее время широкое применение при решении несовместных и неустойчивых систем линейных алгебраических уравнений и неравенств и несобственных задач линейного программирования ([1-5, 7]). Использование методов коррекции в задачах обработки экспериментальных данных в условиях наличия шумов при измерении не только выходных, но и входных переменных, составляет содержание полного метода наименьших квадратов (TLS в англоязычной

терминологии, см. [12-14]). Этот метод имеет вероятностное обоснование как метод максимального правдоподобия в случае нормального распределения ошибок как при измерении векторного аргумента x , так и значений функции y (см. [9]).

Нередко при моделировании случайных процессов в сложных системах, в частности, в экономике и финансах, распределения величин не соответствуют нормальному закону. Это связано с тем, что значение эксцесса больше у статистических распределений, основанных на реальных данных. Такие распределения случайных величин имеют «тяжелые хвосты», т.е. соответствующая плотность распределения медленнее в сравнении с нормальной убывает при $|x| \rightarrow \infty$. Это подтверждается как видом эмпирических плотностей, так и стандартными статистическими критериями. Принято считать, что распределение имеет тяжелый правый хвост, если вероятность того, что случайная величина превосходит достаточно большое x , имеет величину порядка $x^{-\alpha}$ (например, распределения Стюдента, Парето [11]).

Метод наименьших квадратов (в том числе полный) теряет свое вероятностное обоснование, если распределение шумов в статистических данных отличается от нормального. В работе [6] приведены примеры реальных финансовых показателей, которые хорошо описываются двусторонним распределением Лапласа (заметим, что классически под именем Лапласа фигурирует одностороннее экспоненциальное распределение). Это распределение не имеет тяжелый хвост в указанном выше смысле, но все же убывает медленнее по сравнению с нормальным законом. В работе [8] показано, что метод максимального правдоподобия при использовании гипотезы экспоненциального распределения шумов приводит в качестве меры аппроксимации к полиэдральной норме l_1 матрицы (под которой понимается сумма модулей ее элементов, т.е. матрица рассматривается как «расширенный» вектор). В частности, для класса аффинных функций геометрическая интерпретация означает, что в пространстве R^{n+1} требуется найти такую гиперплоскость

$$L: y = (a, x) + b,$$

которая расположена ближе всех гиперплоскостей к заданным точкам в смысле суммы расстояния в метрике l_1 .

Как известно, решение задачи матричной коррекции по норме l_1 сводится к решению ряда задач линейного программирования (различные варианты соответствующих методов содержатся, например, в работах [8, 15, 16]).

В данной работе этот результат обобщается на задачу построения линейной регрессии для данных, которые необходимо предварительно подвергнуть линейному преобразованию. Это приводит к новым более сложным задачам коррекции (аппроксимации) и необходимости разработки методов параметрической коррекции. Показано, что такие задачи тоже можно свести к линейному программированию (ЛП). Разработанные методы проиллюстрированы на тестовых примерах, а затем применены к практической задаче с реальными данными.

Широко распространенной процедурой, предшествующей построению регрессии, является трансформация данных, которая заключается в оптимизации их представлений и форматов (нормализация, стандартизация и т.д.).

Нормализация – это процедура предобработки входных данных, при которой значения признаков (объясняющих переменных) в обучающих или тестовых выборках приводятся к некоторому заданному диапазону (например, $[0,1]$). Нормализация необходима тогда, когда во входном векторе содержатся значения, отличающиеся друг от друга на несколько порядков. Нормализация может быть произведена линейным преобразованием с умножением информационной матрицы X справа на диагональную матрицу, состоящую из различных масштабирующих коэффициентов, и прибавлением матрицы, меняющей начало шкалы отсчета.

Стандартизация – это преобразование данных с гауссовым распределением, которое вычитает среднее значение и делит результат на стандартное отклонение выборки данных. Это приводит к преобразованию данных со средним значением равным нулю, и со стандартным отклонением равным 1. Результирующее распределение называется стандартным распределением

Гаусса или стандартным нормальным. Однако аналогичное преобразование можно применять и к другим законам распределения данных, например, к распределению Лапласа.

С помощью подходящих линейных преобразований можно также изменять свойства многомерных распределений. Так если данные, представленные матрицей X , коррелированы, а Σ есть ковариационная матрица, то для устранения корреляции данных можно использовать разложение $\Sigma = AA^T$ и преобразование $A^{-1}X$, которое приводит к единичной ковариационной матрице.

Значительное внимание в современной литературе по регрессионному анализу уделяется проблеме выбросов – атипичным данным в выборке (см., например, [10]). Такие данные также можно сглаживать с помощью линейных преобразований.

Таким образом, многие (но конечно не все) процедуры трансформации данных могут быть представлены в виде линейного преобразования информационной матрицы, т.е. умножения ее слева или справа на специально подобранную матрицу соответствующей размерности, а также прибавления соответствующей матрицы. Оказывается, что выбор варианта умножения информационной матрицы влияет на способ решения получающейся оптимизационной задачи, поэтому далее они рассматриваются отдельно.

2. Обобщенная задача построения линейной регрессии в метрике l_1 с левосторонним умножением информационной матрицы

Итак, математическая постановка задачи регрессии заключается в следующем. Зависимость величины (числового значения) определенного свойства некоторого процесса или явления от другого свойства зарегистрирована на множестве точек $x^i \in R^n$, $i=1, \dots, m$, множеством значений $y^i \in R$, $i=1, \dots, m$.

Введем в рассмотрение матрицу исходных данных

$$Z = \begin{pmatrix} -y^1 & x_1^1 & \cdots & x_n^1 \\ -y^2 & x_1^2 & \cdots & x_n^2 \\ \cdots & \cdots & \cdots & \cdots \\ -y^m & x_1^m & \cdots & x_n^m \end{pmatrix}.$$

Предполагается, что ошибки наблюдения (шумы) имеют место как во входных данных (объясняющих переменных), так и в выходных данных (объясняемой переменной). Поэтому все элементы этой матрицы будем подвергать линейному преобразованию и аддитивной коррекции. В качестве частных случаев нетрудно рассмотреть варианты коррекции (аппроксимации) только по входным или выходным данным.

В этом разделе мы рассмотрим вариант линейного преобразования исходных данных вида

$$(2) \quad D(Z) = D_0 + D_1 Z,$$

где Z – матрица размера $m \times (n+1)$, D_0 – матрица размера $m \times (n+1)$, D_1 – невырожденная матрица размера $m \times m$.

Использование гипотезы линейной зависимости (объясняемой от объясняющих переменных) предполагает, что после преобразования (2) все точки должны лежать на одной гиперплоскости, т.е. удовлетворять матричному уравнению

$$(3) \quad D(Z)\bar{a} = -be,$$

где $e = (1, 1, \dots, 1)^T \in R^m$, $\bar{a} = (a_0, a_1, \dots, a_n)^T$, $a_0 = 1$, $b \in R$.

Вследствие наличия шумов в исходных данных переопределенная система уравнений (3) относительно неизвестных $\bar{a}$, b является, как правило, несовместной. Задача аппроксимации эквивалентна минимальной коррекции исходных данных, при которой данная система становится совместной.

Обычно в случае необходимости предварительной обработки исходных данных процесс построения регрессии состоит из двух этапов. Сначала производится преобразование исходной информационной матрицы, а затем для полученной информационной матрицы решается задача аппроксимации. Однако гипотеза о наличии шумов с некоторым законом распределения от-

носится, вообще говоря, к исходным данным, полученным экспериментальным путем. Поэтому предлагается идея, состоящая в совместном использовании процедур преобразования и аппроксимации, т.е. в варьировании при аппроксимации исходных данных. Заметим, что наличие свободного члена b в классе аффинных функций означает, по существу, введение дополнительной фиктивной переменной, принимающей значение 1. Поэтому это значение не подлежит коррекции.

Сформулируем постановку задачи коррекции данных, для которой возможна непосредственная коррекция элементов матрицы Z , а элементы матрицы D подвергаются изменению согласно (2).

В качестве минимизируемого критерия величины коррекции здесь будем использовать норму матрицы l_1 . Обозначим матрицу коррекции параметров матрицы Z через H . Эта матрица должна удовлетворять условию $D_0\bar{a} + D_1(Z - H)\bar{a} = -eb$ или

$$(4) \quad H\bar{a} = D_1^{-1}D_0\bar{a} + Z\bar{a} + D_1^{-1}eb.$$

Рассмотрим задачу нахождения минимальной по норме l_1 матрицы H , для которой система уравнений (4) совместна:

$$(5) \quad \min_{\bar{a}, b, H} \{ \|H\|_1 \mid H\bar{a} = D_1^{-1}D_0\bar{a} + Z\bar{a} + D_1^{-1}eb \},$$

где $\|H\|_1$ – норма l_1 матрицы H .

Теорема 1. Задача (5) имеет решение тогда и только тогда, когда существует такое решение (u^*, v^*, w^*, j^*) задачи

$$(6) \quad \min_{u, v, w, j} \{ \langle e, u \rangle \mid u \geq \pm(D_1^{-1}D_0v + Zv + D_1^{-1}ew), \\ u \geq 0, |v_i| \leq 1, i = 0, \dots, n, |v_j| = 1 \},$$

что $v_0^* \neq 0$. Тогда $\bar{a}^* = \frac{v^*}{v_0^*}$, $b^* = \frac{w^*}{v_0^*}$ является решением задачи

(5), элементы матрицы коррекции H^* определяются по формулам

$$(7) \quad h_{ij}^* = \frac{(D_1^{-1}D_0\bar{a}^* + Z\bar{a}^* + D_1^{-1}eb^*)_i}{\max_{0 \leq j \leq n} |\bar{a}_j^*|}, h_{ij}^* = 0, j \neq j^*, i = 1, \dots, m,$$

где $j^* = \arg \max_{0 \leq j \leq n} |\bar{a}_j^*|$, $\|H^*\|_1 = \langle e, u^* \rangle$, $(\dots)_i$ – i -я компонента вектора, стоящего в скобках.

Доказательство. Минимальная по норме l_1 матрица коррекции H при фиксированных $\bar{a}$, b определяется, очевидно, по формулам

$$(8) \quad |h_{ij_0}| = \frac{|(D_1^{-1}D_0\bar{a} + Z\bar{a} + D_1^{-1}eb)_i|}{\max_{0 \leq j \leq n} |\bar{a}_j|}, h_{ij} = 0, j \neq j_0, i = 1, \dots, m,$$

где индекс $j_0 = \arg \max_{0 \leq j \leq n} |\bar{a}_j|$.

Действительно, i -я компонента вектора $H\bar{a}$ левой части формулы (4) представляют собой линейную функцию от элементов i -й строки матрицы H , а коэффициентами являются компоненты вектора $\bar{a}$. Эта линейная функция равна некоторой константе, т.е. удовлетворяет одному ограничению типа равенства. Минимум суммы модулей элементов i -й строки матрицы H достигается, когда все они равны нулю кроме элемента с самым большим по модулю коэффициентом. При этом данный элемент равен отношению константы к модулю соответствующего коэффициента, а его знак определяется знаком константы. Таким образом, при фиксированных $\bar{a}$, b минимальная по норме l_1 матрица коррекции H , удовлетворяющая (4), имеет один ненулевой столбец, модули компонент которого вычисляются по формулам (8). Задача минимальной коррекции сводится к минимизации суммы модулей этих элементов матрицы H по $\bar{a}$, b .

Введем скалярные переменные $r = (\max_{0 \leq j \leq n} |\bar{a}_j^*|)^{-1}$, $w = rb$ и вектор

$v = r\bar{a}$. Компоненты вектора v должны удовлетворять условиям $-1 \leq v_j \leq 1$, причем существует индекс j_0 такой, что $v_{j_0} = 1$ или -1 .

Введем также переменные u_i , которые не меньше выражения в правой части формул (8) и m -мерный вектор $u = (u_1, \dots, u_m)$. В новых переменных получаем задачу математического программирования

$$(9) \min_{u,v,w} \{ \langle e, u \rangle \mid u \geq \pm(D_1^{-1}D_0v + Zv + D_1^{-1}ew), \\ u \geq 0, -1 \leq v_i \leq 1, i = 0, \dots, n, v_j = 1 \vee v_j = -1 \}.$$

Это задача линейного, частично целочисленного программирования. Ее можно свести к решению стандартных задач ЛП. Для этого последовательно полагаем $v_j = \pm 1$ для $j = 0, \dots, n$ и выбираем из получающихся задач ЛП ту, которая дает наименьшее значение критерия $\langle e, u \rangle = \sum_{i=1}^m u_i$. Таким образом, приходим к задаче (6).

По найденному решению этой задачи v^* и w^* с учетом требования $a_0=1$ в случае $v_0^* \neq 0$ находим вектор $\bar{a}^* = \frac{v^*}{v_0^*}$, $b^* = \frac{w^*}{v_0^*}$ и по формулам (7) элементы матрицы коррекции (знак элемента определяется знаком выражения, стоящего внутри модуля). В случае $v_0^* = 0$ задача (5) не имеет решения. Теорема доказана.

Заметим, что в силу однородности ограничений в (9) при значениях $v_j = \pm 1$ решения получаются одинаковыми. Поэтому достаточно решить $n+1$ задачу ЛП.

Рассмотрим иллюстративный пример, демонстрирующий работу предлагаемого метода (для работы с матрицами и решения вспомогательных оптимизационных задач использован математический пакет MatLab).

Пример 1. Пусть исходная информационная матрица есть

$$Z = \begin{pmatrix} -1 & 1 & 10.5 \\ -0.5 & 2 & 20.3 \\ -0.3 & 3 & 15 \\ -0.8 & 2.5 & 15.4 \\ -0.7 & 1 & 20.1 \end{pmatrix}.$$

В качестве матриц линейного преобразования данных возьмем

$$D_0 = \begin{pmatrix} -1 & 1 & 1 \\ -0.5 & 2 & 2 \\ -0.3 & 3 & 3 \\ -0.8 & 2.5 & 2.5 \\ -0.7 & 1 & 1 \end{pmatrix}, D_1 = \begin{pmatrix} 1/\rho & 0 & 0 & 0 & 0 \\ 0 & 1/\rho & 0 & 0 & 0 \\ 0 & 0 & 1/\rho & 0 & 0 \\ 0 & 0 & 0 & 1/\rho & 0 \\ 0 & 0 & 0 & 0 & 1/\rho \end{pmatrix},$$

где $\rho=37.58$ – евклидова норма матрицы Z . Тогда

$$D_0 + D_1 Z = \begin{pmatrix} -1.027 & 1.027 & 1.279 \\ -0.513 & 2.053 & 2.54 \\ -0.308 & 3.08 & 3.399 \\ -0.821 & 2.567 & 2.91 \\ -0.719 & 1.027 & 1.535 \end{pmatrix}.$$

Система уравнений $(D_0 + D_1 Z)\bar{a} = -be$ несовместна, поэтому решим задачу минимальной коррекции (5).

Решая задачу (9) при $v_0=1$, получаем $u=(0, 0, 3.526, 11.419, 1.974)^T$, $v=(1, 0.728, -1)^T$, $w=1.559$, $\langle e, u \rangle=16.919$, $\bar{a}=(1, 0.728, -1)^T$, $b=1.559$.

Решая задачу (9) при $v_1=1$, получаем $u=(0, 0.3, 9.4, 0, 4.5)^T$, $v=(-0.324, -0.914, -1)^T$, $w=-0.829$, $\langle e, u \rangle=14.2$, $\bar{a}=(1, 2.269, -2.269)^T$, $b=1.6$.

Решая задачу (9) при $v_2=1$, получаем $u=(5.85, 0, 5.5, 0, 0)^T$, $v=(0.441, 1, -1)^T$, $w=0.705$, $\langle e, u \rangle=11.35$, $\bar{a}=(1, 2.822, -3.086)^T$, $b=2.558$.

Третий случай дает меньшее значение $\langle e, u \rangle=11.35$ задачи (9), следовательно, $\bar{a}^*=(1, 2.822, -3.086)^T$, $b^*=2.558$ – решение

задачи (5) с матрицей коррекции $H^* = \begin{pmatrix} 0 & 0 & -5.85 \\ 0 & 0 & 0 \\ 0 & 0 & -5.5 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. При этом

$$Z - H = \begin{pmatrix} -1 & 1 & 16.35 \\ -0.5 & 2 & 20.3 \\ -0.3 & 3 & 20.5 \\ -0.8 & 2.5 & 15.4 \\ -0.7 & 1 & 20.1 \end{pmatrix}. \text{ После коррекции положения точек}$$

по критерию минимума в метрике l_1 они принадлежат одной плоскости, уравнение которой имеет вид

$$y = 2.822x_1 - 3.086x_2 + 2.558.$$

3. Обобщенная задача построения линейной регрессии в метрике l_1 с правосторонним умножением информационной матрицы

В данном разделе мы рассмотрим вариант линейного преобразования исходных данных вида

$$(10) \quad D(Z) = D_0 + ZD_2,$$

где Z – матрица размера $m \times (n+1)$, D_0 – матрица размера $m \times (n+1)$, D_2 – невырожденная матрица размера $(n+1) \times (n+1)$.

Гипотеза линейной зависимости (объясняемой от объясняющих переменных) предполагает, что после преобразования (10) все точки должны лежать на одной гиперплоскости, т.е. удовлетворять матричному уравнению (3).

Сформулируем постановку задачи коррекции данных, для которой возможна непосредственная коррекция элементов матрицы Z , а элементы матрицы D подвергаются изменению согласно (10).

В качестве минимизируемого критерия величины коррекции здесь будем также использовать норму матрицы l_1 . Матрица H коррекции параметров матрицы Z должна удовлетворять условию $D_0\bar{a} + (Z - H)D_2\bar{a} = -eb$. В данном случае использованный в первом разделе прием умножения на обратную матрицу не проходит. Поэтому здесь мы используем другой прием, а именно, замену переменных $\tilde{a} = D_2\bar{a}$. Если мы найдем вектор

$\tilde{a}$, то коэффициенты регрессии определяются по формуле $\bar{a} = D_2^{-1} \tilde{a}$. В новых переменных получаем уравнение

$$(11) \quad H\tilde{a} = D_0 D_2^{-1} \tilde{a} + Z\tilde{a} + eb.$$

Рассмотрим задачу нахождения минимальной по норме l_1 матрицы H , для которой система уравнений (11) совместна:

$$(12) \quad \min_{\tilde{a}, b, H} \{ \|H\|_1 \mid H\tilde{a} = D_0 D_2^{-1} \tilde{a} + Z\tilde{a} + eb \}.$$

Теорема 2. Задача (12) имеет решение тогда и только тогда, когда существует такое решение (u^*, v^*, w^*, j^*) задачи

$$(13) \quad \min_{u, v, w, j} \{ \langle e, u \rangle \mid u \geq \pm(D_0 D_2^{-1} v + Zv + ew), \\ u \geq 0, |v_i| \leq 1, i = 0, \dots, n, |v_j| = 1 \},$$

что $v_0^* \neq 0$. Тогда $\tilde{a}^* = \frac{v^*}{v_0^*}$, $b^* = \frac{w^*}{v_0^*}$ является решением задачи

(13), элементы матрицы коррекции H^* определяются по формулам

$$(14) \quad h_{ij}^* = \frac{(D_0 D_2^{-1} \tilde{a}^* + Z\tilde{a}^* + eb^*)_i}{\max_{0 \leq j \leq n} |\tilde{a}_j^*|}, h_{ij}^* = 0, j \neq j^*, i = 1, \dots, m,$$

где $j^* = \arg \max_{0 \leq j \leq n} |\tilde{a}_j^*|$, $\|H^*\|_1 = \langle e, u^* \rangle$, $(\dots)_i$ — i -я компонента вектора, стоящего в скобках.

Доказательство теоремы 2 аналогично доказательству теоремы 1, мы его приводить не будем.

Исходная задача опять сводится к решению $n+1$ задач ЛП и последующему определению коэффициентов регрессии по формуле $\bar{a} = D_2^{-1} \tilde{a}$.

Рассмотрим пример решения задачи аппроксимации с преобразованием умножения информационной матрицы справа на некоторую матрицу.

Пример 2. Возьмем исходную информационную матрицу Z из примера 1, а в качестве матриц линейного преобразования данных

$$D_0 = (0)_{5 \times 3}, \quad D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/\alpha \end{pmatrix},$$

где $\alpha = \max \{10.5, 20.3, 15, 15.4, 20.1\} = 20.3$. Тогда

$$D_0 + ZD_2 = \begin{pmatrix} -1 & 1 & 0.517 \\ -0.5 & 2 & 1 \\ -0.3 & 3 & 0.739 \\ -0.8 & 2.5 & 0.759 \\ -0.7 & 1 & 0.99 \end{pmatrix}.$$

Система уравнений $(D_0 + ZD_2)\bar{a} = -eb$ несовместна, поэтому решаем задачу минимальной коррекции (12).

Решая задачу (13) при $v_0=1$, получаем $u=(0, 0, 0.172, 0.244, 0)^T$, $v=(1, -0.194, -0.031)^T$, $w=1.522$, $\langle e, u \rangle=0.416$, $\tilde{a}=(1, -0.194, -0.031)^T$, $b=1.522$.

Решая задачу (13) при $v_1=1$, получаем $u=(1.979, 0, 0, 0.06, 0.83)^T$, $v=(-1, 1, 0.151)^T$, $w=-5.564$, $\langle e, u \rangle=2.870$, $\tilde{a}=(1, -1, -0.151)^T$, $b=5.564$.

Решая задачу (13) при $v_2=1$, получаем $u=(6.2, 4.1, 0.4, 0, 3.1)^T$, $v=(-1, 1, 1)^T$, $w=-18.7$, $\langle e, u \rangle=13.8$, $\tilde{a}=(1, -1, -1)^T$, $b=18.7$.

Первый случай дает меньшее значение $\langle e, u \rangle=0.416$ задачи (13), следовательно, $\tilde{a}^*=(1, -0.194, -0.031)^T$, $b^*=1.522$ есть ре-

шение задачи (12) с матрицей коррекции $H^* = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0.172 \\ 0 & 0 & -0.244 \\ 0 & 0 & 0 \end{pmatrix}.$

$$\text{При этом } Z - H = \begin{pmatrix} -1 & 1 & 10.5 \\ -0.5 & 2 & 20.3 \\ -0.3 & 3 & 14.828 \\ -0.8 & 2.5 & 15.644 \\ -0.7 & 1 & 20.1 \end{pmatrix}. \text{ Коэффициенты регрес-}$$

сии есть $\bar{a}^* = (1, -0.194, -0.634)^T$. После коррекции положения точек по критерию минимума в метрике l_1 они принадлежат одной плоскости, уравнение которой имеет вид

$$y = -0.194x_1 - 0.634x_2 + 1.522.$$

4. Пример преобразования данных при построении демографического тренда

Рассмотрим практический пример, основанный на реальных данных, взятых с сайта Росстат [17]. Мы будем анализировать показатель рождаемости в Федеральных округах в зависимости от двух факторов: численности городского населения и доходов.

В таблице 1 приведены данные за 2018 год по Федеральным округам РФ: коэффициент рождаемости, удельный вес городского населения, доходы населения.

Анализ показателей указывает, что Северо-Кавказский регион выделяется среди остальных регионов, он имеет один из самых низких доходов на душу населения, самый низкий процент городского населения, а коэффициент рождаемости значительно выше значений этого показателя в остальных регионах. Таким образом, можно предположить, что в данном случае мы имеем дело с наличием выброса. С одной стороны, этот регион является элементом общей картины и должен учитываться при построении демографического тренда, с другой стороны он является весьма атипичным для РФ. По-видимому, здесь рождаемость определяется другими важными факторами (традициями, религией, большим подсобным хозяйством и т.п.).

Таблица 1. Анализируемые социально-экономические показатели по регионам России за 2018 год

Федеральные округа	Общий коэффициент рождаемости (число родившихся на 1000 чел. населения)	Удельный вес городского населения (в %)	Среднедушевые доходы (в месяц)
Центральный	10.4	82.2	40.843
Северо-Западный	11.1	84.4	33.89
Южный	11.1	62.6	26.928
Северо-Кавказский	15	49.8	24.017
Приволжский	11.1	71.9	25.87
Уральский	12.6	81.4	32.944
Сибирский	12.3	73.1	23.925
Дальневосточный	12.1	75.8	37.07

По исходным данным построим линейную регрессию в метрике l_1 без преобразования данных, т.е. $D_0=(0)_{8 \times 3}$ (нулевая матрица), а $D_1=E_{8 \times 8}$ (единичная матрица). Минимальное значение нормы матрицы коррекции $\langle e, u \rangle = 5.997$. Уравнение регрессии имеет вид $y = -0.116x_1 + 0.006x_2 + 20.563$.

Исключим Северо-Кавказский федеральный округ и построим линейную регрессию в метрике l_1 без преобразования данных, т.е. $D_0=(0)_{7 \times 3}$, а $D_1=E_{7 \times 7}$. Минимальное значение нормы матрицы коррекции $\langle e, u \rangle = 3.980$. Уравнение регрессии имеет вид $y = 0.7106x_1 - 0.151x_2 + 10.698$.

Таким образом, включение этого региона полностью изменяет результат: коэффициенты регрессии меняют знаки. Это подтверждает предварительное заключение о наличии выброса.

По исходным данным построим линейную регрессию в метрике l_1 с левосторонним умножением информационной матрицы. В качестве матриц линейного преобразования данных возьмем

$$D_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ -10 & 49.8 & 20.017 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, D_1 = \begin{pmatrix} 1 & 0 & 0 & 1/\rho & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1/\rho & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1/\rho & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/\rho & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/\rho & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/\rho & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/\rho & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1/\rho & 0 & 0 & 0 & 1 \end{pmatrix},$$

где $\rho=1/84.4$. Тогда $D_0 + D_1 Z =$

$$\begin{pmatrix} -10.58 & 82.79 & 41.13 \\ -11.28 & 84.99 & 34.17 \\ -11.28 & 63.19 & 27.21 \\ -10.18 & 50.39 & 24.30 \\ -11.28 & 72.49 & 25.15 \\ -12.78 & 81.99 & 33.23 \\ -12.48 & 73.69 & 24.21 \\ -12.28 & 76.39 & 37.35 \end{pmatrix}.$$

Минимальное значение нормы матрицы коррекции $\langle e, u \rangle = 4.429$. Уравнение регрессии имеет вид $y = 0.098x_1 - 0.165x_2 + 9.247$.

По исходным данным построим линейную регрессию в метрике l_1 с правосторонним умножением информационной матрицы. В качестве матриц линейного преобразования данных возьмем

$$D_2 = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$D_0 = \begin{pmatrix} 100 & 0 & 0 \\ 100 & 0 & 0 \\ 100 & 0 & 0 \\ 146 & 0 & 0 \\ 100 & 0 & 0 \\ 100 & 0 & 0 \\ 100 & 0 & 0 \\ 100 & 0 & 0 \end{pmatrix}. \text{ Тогда } D_0 + ZD_2 = \begin{pmatrix} -4 & 82.2 & 40.84 \\ -11 & 84.4 & 33.89 \\ -11 & 62.6 & 26.93 \\ -4 & 49.8 & 24.02 \\ -11 & 71.9 & 25.87 \\ -26 & 81.4 & 35.94 \\ -23 & 73.1 & 23.93 \\ -21 & 75.8 & 37.07 \end{pmatrix}.$$

Минимальное значение нормы матрицы коррекции $\langle e, u \rangle = 4.143$. Уравнение регрессии имеет вид $y = x_1 - 1.926x_2 + 0.113$.

Результаты расчетов показывают, что сглаживание данных позволяет сохранить общий вид тренда и, в то же время, улучшить значение критерия аппроксимации, тогда как полное исключение точки выброса в корне меняет вид тренда. Конечно, выбор линейного преобразования носит эмпирический характер, что затрудняет объективное сравнение с другими подходами, которые также, по существу, носят эвристический характер (см., например, [10]). По существу, здесь требуется найти компромисс между требованиями устойчивости общего вида тренда и уменьшения меры аппроксимации.

5. Заключение

В данной статье была рассмотрена несобственная задача интерполяции и подход к ее решению, основанный на линейном преобразовании информационной матрицы и матричной коррекции получающейся системы линейных уравнений.

В качестве меры аппроксимации использована полиэдральная норма матрицы l_1 , которая имеет вероятностное обоснование как метод максимального правдоподобия при двухстороннем экспоненциальном распределении шумов во входных и выходных данных и приводит к решению задач ЛП. Заметим, что данный подход в вычислительном плане оказывается проще по

сравнению полным методом наименьших квадратов, который приводит к более сложным вычислительным задачам нахождения собственных чисел и векторов матриц большой размерности даже без преобразования данных (как совместить его с преобразованием данных вообще пока неясно).

Заметим, что вариант преобразования данных с правосторонним умножением на матрицу позволяет нормировать каждый столбец исходной информационной матрицы в отдельности, т.е. удобен при соизмерении данные с сильно различающимися диапазонами значений, а вариант преобразования данных с левосторонним умножением на матрицу позволяет нормировать каждую строку исходной информационной матрицы в отдельности, т.е. решать проблему выбросов.

Литература

1. БАРКАЛОВА О.С. *Коррекция несобственных задач линейного программирования в канонической форме по минимаксному критерию* // Журнал вычислительной математики и математической физики. – 2012. – Т.52. – №12. – С. 2178–2189.
2. ВАТОЛИН А.А. *Аппроксимация несобственных задач линейного программирования по критерию евклидовой нормы* // Журн. вычисл. матем. и матем. физика. – 1984. – Т.24. – № 12. – С. 1907–1908.
3. 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.
4. ГОРЕЛИК В.А., МУРАВЬЕВА О.В. *Методы коррекции несобственных задач и их применение к проблемам оптимизации и классификации*. – М.: ВЦ РАН, 2012. – 148 с.
5. ГОРЕЛИК В.А., ЕРОХИН В.И., ПЕЧЕНКИН Р.В. *Численные методы коррекции несобственных задач линейного программирования и структурных систем уравнений*. – М.: ВЦ РАН, 2006. – 150 с.

6. ГОРЕЛИК В.А., ЗОЛотова Т.В. *Вопросы формирования оптимального портфеля акций российских компаний с вероятностной функцией риска* // Научно-исследовательский финансовый институт. Финансовый журнал. – 2016. – №3. – С. 35–44.
7. ЕРОХИН В.И., КРАСНИКОВ А.С., ХВОСТОВ М.Н. *Минимальные по евклидовой норме матричные коррекции задач линейного программирования* // Автоматика и телемеханика. – 2012. – №2. – С. 11–24.
8. ГОРЕЛИК В.А., ТРЕМБАЧЕВА (БАРКАЛОВА) 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.
9. МАТРОСОВ В.Л., ГОРЕЛИК В.А., ЖДАНОВ С.А., МУРАВЬЕВА О.В., УГОЛЬНИКОВА Б.З. *Теоретические основы информатики: учебное пособие для студентов вузов*. М.: Издательский центр «Академия», 2018. – 352 с.
10. ШИБЗУКОВ З.М. *О принципе минимизации эмпирического риска на основе усредняющих агрегирующих функций* // Доклады Академии наук. – 2017. – Т. 476. – № 5. – С. 495-499.
11. ШИРЯЕВ А.Н. *Основы стохастической финансовой математики. Т. 1. Факты. Модели*. М.: ФАЗИС, 1998. – 512 с.
12. BACK A. *The Matrix-Restricted Total Least Squares Problem* // Signal Process. – 2007. – No. 87(10). – P. 2303-2312.
13. MARKOVSKY I. *Bibliography on total least squares and related methods* // Statistics and its interface. – 2010. – Vol. 2. – P. 1–6.
14. MARKOVSKY I., VAN HUFFEL S. *Overview on total least squares methods* // Signal Processing. – 2007. – Vol. 87. – P. 2283–2302.
15. OSBORNE M., WATSON J. *An analysis of the total approximation problem in separable norms, and an algorithm for the total l_1 problem.* // SIAM J. Sci. Stat. Comput. – 1985. – Vol. 6. – No. 2. – P. 410–424.

16. ROSEN J. B., PARK H., GLICK J. *Total least norm formulation and solution for structured problems* // SIAM Journal on Matrix Anal. Appl. – 1996. – Vol. 17. – No. 1. – P. 110–128.
17. Регионы России. Социально-экономические показатели 2018. Режим доступа
<https://www.gks.ru/storage/mediabank/reg-pok%2018.pdf>

THE PROBLEM OF A LINEAR REGRESSION CONSTRUCTION WITH THE DATA TRANSFORMATION AND L_1 METRIC USING

Victor Gorelik, Dorodnicyn Computing Centre FRC CSC RAS, Moscow, Doctor of Science, professor (vgor16@mail.ru).

Tatiana Zolotova, Financial University under the Government of the Russian Federation, Moscow, Doctor of Science, professor (tgold11@mail.com).

Abstract: The linear regression analysis problem is considered under the assumption of the presence of noise in the output and input variables, i.e. the problem of constructing an affine function, which in a certain sense describes best the original data. The geometric interpretation consists in constructing a hyper plane in $(n+1)$ -dimensional space, which runs the closest to the given points in the sense of the chosen measure of proximity. This approximation problem can also be interpreted as an improper (has no solution) interpolation problem, for which it is required to change (correct) optimally the positions of the points so that they all lie on the same hyper plane. To estimate the measure of approximation (correction) of the initial data, the l_1 metric is used, which has a probabilistic rationale as a maximum likelihood method for two-sided exponential noise distribution. In addition, the case of preliminary linear data transformation is considered, while the original information matrix is subject to correction. Therefore, the arising problems belong to the new class of parametric correction. It is shown that these approximation (corrections) problems can be reduced to a set of a finite number of linear programming problems. The MatLab math package was used to work with matrices and solve auxiliary optimization problems.

Keywords: data processing, regression analysis, matrix correction, maximum likelihood method, exponential distribution.

УДК 519.7

ББК 22.18

*Статья представлена к публикации
членом редакционной коллегии ...заполняется редактором...*

Поступила в редакцию ...заполняется редактором...

Опубликована ...заполняется редактором...