УДК 519.237.5:519.24 DOI: 10.14529/mmph180205

ТОЧНОЕ ОЦЕНИВАНИЕ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ МЕТОДОМ НАИМЕНЬШИХ МОДУЛЕЙ НА ОСНОВЕ СПУСКА ПО УЗЛОВЫМ ПРЯМЫМ

А.Н. Тырсин[1,2]***, А.А. Азарян***[2]

[1] *Научно-инженерный центр «Надежность и ресурс больших систем и машин» УрО РАН, г. Екатеринбург, Российская Федерация*

[2] *Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, г. Екатеринбург, Российская Федерация*

E-mail: at2001@yandex.ru

При построении линейных моделей во многих случаях приходится сталкиваться со стохастической неоднородностью экспериментальных данных. Это проявляется в нарушении условий теоремы Гаусса–Маркова, в частности наблюдения могут быть засорены грубыми ошибками. В этих условиях оценивание параметров моделей требуется выполнять с помощью устойчивых методов. К их числу относят метод наименьших модулей. Однако известные алгоритмы его реализации являются достаточно эффективными лишь для малых размерностей моделей и ограниченного объема выборок. Цель данного исследования – разработка эффективных вычислительных алгоритмов реализации метода наименьших модулей, не имеющих ограничений на порядок моделей и объем экспериментальных данных. Описаны алгоритмы точного решения задачи оценивания параметров линейных регрессионных моделей методом наименьших модулей. Они основаны на спуске по узловым прямым. Для снижения вычислительных затрат использована особенность узловых прямых – все расположенные на каждой такой прямой узловые точки являются пересечением набора гиперплоскостей, из которых различными является только одна гиперплоскость. Данные алгоритмы значительно выигрывают по сравнению с известным переборным алгоритмом и могут эффективно использоваться на практике. Получена оценка вычислительной сложности алгоритма спуска по узловым прямым. Приведена схема алгоритма.

Ключевые слова: метод наименьших модулей; линейная регрессионная модель; алгоритм; узловая точка; узловая прямая; гиперплоскость; вычислительная сложность.

Введение

Одной из распространенных задач при статистической обработке результатов экспериментальных исследований является оценивание неизвестных коэффициентов линейной регрессионной модели [1–4]

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \boldsymbol{\varepsilon}, \tag{1}$$

где $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} \in \mathbf{R}^n$ – вектор измерений; $\mathbf{X} = \{x_{ij}\}_{n\times m} = \begin{pmatrix} 1 & x_{12} & \dots & x_{1m} \\ 1 & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n2} & \dots & x_{nm} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdots \\ \mathbf{x}_n \end{pmatrix}$ – заданная матрица; $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix} \in \mathbf{R}^n$ – вектор случайных ошибок; $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_m \end{pmatrix} \in \mathbf{R}^m$ – вектор коэффициентов регрессии.

Среди прикладных статистических методов построения регрессионных зависимостей наибольшее распространение получил метод наименьших квадратов (МНК). Использование МНК требует выполнения ряда предпосылок, называемых условиями Гаусса–Маркова [4, с. 6–7]. При их выполнении МНК-оценки параметров модели (1) являются состоятельными и несмещенными. Кроме того, если случайные ошибки имеют нормальный закон распределения, то МНК-оценки становятся эффективными. Однако использование МНК при нарушении условий Гаусса–Маркова может привести к значительным ошибкам при оценивании коэффициентов, а в случае присутствия в измерениях больших выбросов даже к несостоятельности оценок.

С целью обеспечения устойчивости оценок относительно отклонений случайных ошибок от гауссовой модели разработан ряд статистических методов, основанных на более общих предположениях относительно случайных ошибок. Одним из них является метод наименьших модулей (МНМ) [5, 6]. МНМ для задачи (1) имеет вид

$$Q(\mathbf{a})=\sum_{i=1}^{n}\left|y_i-\langle\mathbf{a},\mathbf{x}_i\rangle\right|\to\min_{\mathbf{a}\in\mathbf{R}^m}. \quad (2)$$

Целевая функция $Q(\mathbf{a})$ задачи (2) является выпуклой. Поскольку во всех точках, где у функции $Q(\mathbf{a})$ существует производная, она линейна, то, следовательно, ее локальные минимумы могут быть только в особых точках. Известно несколько точных и приближенных методов решения задачи (2) [6–9].

Кусочно-линейный вид целевой функции позволяет рассчитывать на наличие быстрых и точных методов решения задачи (2). Однако в настоящее время в целом удовлетворительно решена только проблема ее приближенного решения. Здесь можно отметить метод вариационно-взвешенных квадратических приближений (называемый также алгоритмом Вейсфельда) [6, 7], подход, основанный на сведении данной проблемы к задаче линейного программирования и ее приближенное решение методом внутренней точки [10], численные методы спуска нулевого порядка [9]. Недостатком этих методов в первую очередь является неточное решение при относительной простоте целевой функции.

К точным методам относится сведение (2) к эквивалентной задаче линейного программирования [6, 11] и ее решение при помощи широко известного симплекс-метода и переборный алгоритм [9].

Использование симплекс-метода ограничено задачами малой размерности. Это вызвано накоплением погрешностей из-за ошибок округлений и требованием чрезмерно большого выделения памяти. Предложенное в [12] использование массивно-параллельных вычислений с использованием дробно-рациональных вычислений без округления устраняет проблему накопления вычислительных погрешностей, но приводит к существенному усложнению реализации и росту вычислительных затрат, затрудняющих широкое практическое применение данного подхода.

Второй метод [9] основан на переборе всех узловых точек, в которых не существует производная функции $Q(\mathbf{a})$ ни по одному из возможных направлений пространства $\mathbf{R}^m$. Переборный алгоритм требует решения C_n^m систем линейных уравнений порядка m. Вычислительные погрешности здесь незначительны и не накапливаются. Однако с ростом n и m наблюдается экспоненциальный рост вычислительных затрат. Фактически практическое применение переборного алгоритма ограничено объемом выборки $n<200$ и числом коэффициентов регрессии $m<4$. Однако отсутствие эффекта накопления вычислительных погрешностей делает возможным использование данного метода при условии ограничения числа перебираемых узловых точек.

Целью статьи является описание эффективного, точного вычислительного метода решения задачи (2), осуществляющего неполный перебор особых точек.

Алгоритмы спуска по узловым прямым

Обычный спуск. Введем гиперплоскости $\Omega_i=\Omega(\mathbf{a},\mathbf{x}_i,y_i)$ в виде уравнений:

$$y_i-\langle\mathbf{a},\mathbf{x}_i\rangle=0,\ (i=1,2,\ldots,n). \quad (3)$$

Зададим также узловые точки пересечения гиперплоскостей (3):

$$\mathbf{u}_{(k_1,\ldots,k_m)}=\bigcap_{s\in M}\Omega_s,\ M=\{k_1,\ldots,k_m\},\ k_1<k_2<\ldots<k_m,\ k_l\in\{1,2,\ldots,n\}. \quad (4)$$

Обозначим U – множество всех узловых точек (4). Алгоритм точного решения задачи (2) основан на спуске к точному решению, двигаясь вдоль узловых прямых l, каждая из которых является пересечением $(m-1)$ различных гиперплоскостей Ω_i:

$$l_{(k_1,\ldots,k_{m-1})}:\bigcap_{i=k_1}^{k_{m-1}}\Omega_i\,,\; k_l\in\{1,2,\ldots,n\}\,. \qquad (5)$$

В качестве начального приближения берется произвольная узловая точка, являющаяся пересечением m гиперплоскостей $\Omega_{k_1},\ldots,\Omega_{k_m}$. Исключив одну из гиперплоскостей, получим узловую прямую l. В любой узловой точке можно построить m таких узловых прямых. Выберем ту, вдоль которой целевая функция достигает наименьшего значения, которое всегда будет достигаться в одной из узловых точек. Найдя эту точку, продолжим движение из нее по тому же принципу. В результате будет найдена узловая точка, спуск из которой невозможен. И эта узловая точка будет являться точным решением задачи (2). В основе алгоритма лежат следующие теоремы.

Теорема 1. Рассмотрим модель (1), для которой имеется выборка наблюдений $(\mathbf{x}_i, y_i)=(x_{i1},x_{i2},\ldots,x_{im},y_i)$, $(i = 1, \ldots , n)$, и пусть заданы функция $Q(\mathbf{a})$ задачи (2), гиперплоскости (3) и множество U всех узловых точек (4). Тогда функция (2) всегда имеет точку глобального минимума, эта точка либо единственна и принадлежит U, либо состоит из выпуклого линейного многогранника, вершины которого являются точками из U.

Доказательство теоремы 1. Известно, что выпуклая, непрерывная, кусочно-линейная функция либо имеет глобальный минимум, либо стремится к минус бесконечности. А поскольку функция $Q(\mathbf{a})$ является также ограниченной снизу ($Q(\mathbf{a})\ge 0$ как сумма модулей) функцией, то она всегда имеет точку глобального минимума.

Пусть $\mathbf{a}^*=(a_1^*,a_2^*,\ldots,a_m^*)^T$ – стационарная точка и является точкой минимума функции $Q(\mathbf{a})$. Тогда ее градиент в этой точке равен нулю. Поскольку $Q(\mathbf{a})$ кусочно-линейная функция, то из равенства $\operatorname{grad}Q(\mathbf{a}^*)=\mathbf{0}$ следует, что функция $Q(\mathbf{a})$ является постоянной функцией на выпуклом многограннике с вершинами $\mathbf{a}^1,\mathbf{a}^2,\ldots,\mathbf{a}^l$, гранями которой являются гиперплоскости (3), имеющем вид [13]:

$$A=\{\mathbf{a}:\mathbf{a}=\sum_{k=1}^{l}\lambda_k^2\mathbf{a}^k,\ \sum_{k=1}^{l}\lambda_k^2=1\}\,.$$

Функция $Q(\mathbf{a})$ является постоянной до граничных точек многогранника A, лежащих на гиперплоскостях (3) Следовательно, вершины $\mathbf{a}^1,\mathbf{a}^2,\ldots,\mathbf{a}^l$ этого многогранника являются узловыми точками и функция $Q(\mathbf{a})$ достигает минимума и в этих точках.

Все гиперплоскости (3), и только они являются особыми точками функции $Q(\mathbf{a})$, поскольку только в них она не дифференцируема. Если взять $m-1$ произвольных невырожденных соотношений вида $l_i=y_i-\langle\mathbf{a},\mathbf{x}_i\rangle=0$, то они в совокупности определяют прямую (5) в пространстве $(a_1,a_2,\ldots,a_m)$ и вместе с тем плоскость $P_{(Q,k_1,\ldots,k_{m-1})}$, параллельную оси Q в пространстве $(Q,a_1,a_2,\ldots,a_m)$. Присоединяя к системе (5) выражение (2) и рассматривая их совместно, найдем уравнение ломанной M, полученной в результате пересечения поверхности (2) плоскостью $P_{(Q,k_1,\ldots,k_{m-1})}$.

Если с помощью уравнений, входящих в систему (5), выразить $m-1$ неизвестных $a_1,a_2,\ldots,a_{m-1}$ через оставшееся неизвестное и подставить в выражение (2), то получим уравнение проекции ломанной M на плоскость (Q, a_m) (рис. 1). Точки $T_{i_1},T_{i_2},\ldots,T_{i_{n-m+1}}$ являются проекциями на эту плоскость точек пересечения прямой $l_{(k_1,\ldots,k_{m-1})}$ гиперплоскостями (3), не вошедшими в (5).

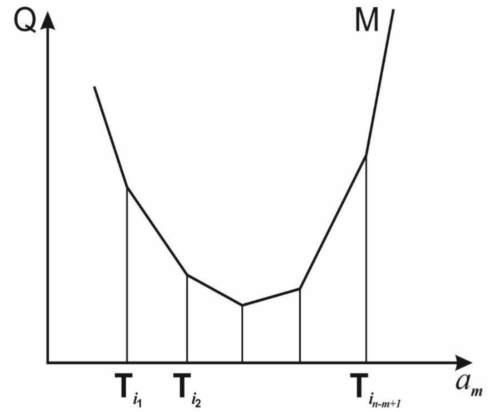

Рис. 1. Вид функции

Функция $Q(a_m)$ на плоскости (Q, a_m) – выпуклая и кусочно-линейная с особыми точками $T_{i_1}, T_{i_2}, \ldots, T_{i_{n-m+1}}$, которые являются узловыми точками. Отсюда следует, что функция $Q(\mathbf{a})$ достигает минимума в узловой точке.

Теорема 2. Алгоритм спуска вдоль узловых прямых для нахождения решения задачи (2), сходится к точному решению за конечное число шагов.

Доказательство теоремы 2. На каждом шаге алгоритма мы находим узловую точку с все меньшим значением целевой функции, а поскольку количество узловых точек конечно, то алгоритм выполняется за конечное число шагов.

Докажем, что алгоритм всегда будет достигать минимума целевой функции (т. е. решения). Допустим, что мы находимся на k-м шаге алгоритма в какой-то узловой точке и пытаемся осуществить спуск вдоль одной из m узловых прямых, проходящих через эту точку, к другой узловой точке с меньшим значением целевой функции. Покажем, что либо мы сможем найти такую точку, либо текущая узловая точка является минимальной.

Целевая функция $Q(\mathbf{a})$ является выпуклой функцией в пространстве $\mathbf{R}^m$. Поэтому, если взять произвольную точку $\tilde{\mathbf{x}}$ и m линейно независимых прямых $\tilde{l}_j$, $(j = 1, 2, \ldots, m)$, проходящих через нее, то либо найдется такая окрестность этой точки, в которой целевая функция будет убывать хотя бы вдоль одного из направлений, либо в текущей узловой точке достигается минимум целевой функции.

Поскольку вероятность того, что случайная величина (случайные погрешности) более одного раза примет одно и то же значение, равна нулю, то узловые прямые, проходящие через узловую точку, являются линейно независимыми. То есть вдоль одной из них целевая функция убывает, причем она будет убывать вплоть до следующей узловой точки на этой узловой прямой, поскольку между двумя соседними узловыми точками, целевая функция (2) будет линейна, так как все ее подмодульные выражения не будут менять знак. Действительно, смена знака означает пересечение одной из образующих гиперплоскостей, что приведет к образованию узловой точки, а мы рассматриваем часть узловой прямой между двумя соседними узловыми точками.

На практике из-за конечной точности измерений гипотетически могут возникнуть ситуации, когда узловые прямые, проходящие через узловую, окажутся параллельными. В этом случае узловой точки не существует, и переходим к рассмотрению другой узловой прямой.

Таким образом, либо произвольная узловая точка является минимумом целевой функции, либо хотя бы по одной из проходящих через нее узловых прямых можно сделать переход к узловой точке с меньшим значением целевой функции. Теорема доказана.

Поясним этот алгоритм, используя рис. 1 (для случая m = 2). В качестве начальной точки выберем точку H. Через нее проходят прямые I и II. Сначала рассмотрим прямую I. Среди узловых точек, которые лежат на этой прямой, выбираем ту, в которой достигается минимальное значение целевой функции (таким образом, происходит спуск по прямой). Предположим, что такой точкой является точка C. Рассмотрим прямую II. Спускаясь по прямой II, находим на ней точку, в которой достигается минимальное значение. Пусть, например, эта точка L. Далее, сравнивая значения целевой функции в точках C и L, выбираем ту, в которой достигается минимальное значение. Пусть, например, эта точка C. Через нее помимо прямой I проходит прямая III. На этой прямой находим очередную точку, в которой достигается минимальное значение целевой функции. Далее, сравнивая значения целевой функции в точках C и D, выбираем ту, в которой достигается минимальное значение. Пусть такой точкой является D.

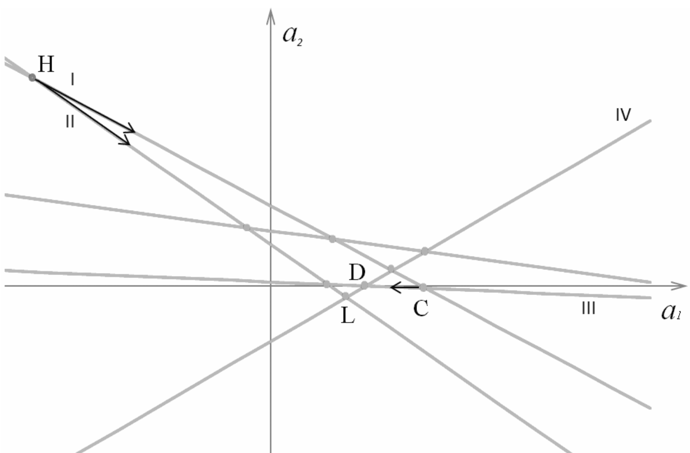

Рис. 2. Спуск по узловым прямым

Допустим, что на прямой IV, проходящей через D, точкой минимума является сама точка D. Тогда в качестве решения выбираем точку D, в которой достигается минимум целевой функции (если бы нашлась другая точка минимума на этой прямой, то из нее продолжили бы дальнейший спуск).

Спуск с использованием разреженных матриц. Двигаясь вдоль прямой $l_{(k_1,\ldots,k_{m-1})}$, для нахождения узловых точек, принадлежащих этой прямой, нужно для каждой точки решать систему линейных алгебраических уравнений (СЛАУ) порядка m:

$$\begin{cases} a_1 + a_2 x_{k_1,2} + a_3 x_{k_1,3} + \ldots + a_m x_{k_1,m} = y_{k_1}, \\ a_1 + a_2 x_{k_2,2} + a_3 x_{k_2,3} + \ldots + a_m x_{k_2,m} = y_{k_2}, \\ \ldots \\ a_1 + a_2 x_{k_{m-1},2} + a_3 x_{k_{m-1},3} + \ldots + a_m x_{k_{m-1},m} = y_{k_{m-1}}, \\ a_1 + a_2 x_{i,2} + a_3 x_{i,3} + \ldots + a_m x_{i,m} = y_i, \end{cases} \tag{6}$$

где $k_1 < k_2 < \ldots < k_{m-1}$, $i \in \{1, 2, \ldots, n\}$, $i \notin \{k_1, k_2, \ldots, k_{m-1}\}$.

Очевидно, что СЛАУ двух различных узловых точек, принадлежащих этой прямой, отличаются лишь одним (последним) уравнением. Следовательно, вычислительная эффективность алгоритма спуска существенно повысится, если для нахождения узловых точек, которые лежат на прямой $l_{(k_1,\ldots,k_{m-1})}$, первые $(m-1)$ строк расширенной матрицы, соответствующей СЛАУ (6), предварительно преобразуем с помощью элементарных преобразований к ступенчатому виду.

Расширенная матрица СЛУ прямой $l_{(k_1,\ldots,k_{m-1})}$ имеет вид

$$\mathrm{A}_{(k_1,\ldots,k_{m-1})} = \begin{pmatrix} 1 & x_{k_1,2} & x_{k_1,3} & \cdots & x_{k_1,m-1} & x_{k_1,m} & y_{k_1} \\ 1 & x_{k_2,2} & x_{k_2,3} & \cdots & x_{k_2,m-1} & x_{k_2,m} & y_{k_2} \\ 1 & x_{k_3,2} & x_{k_3,3} & \cdots & x_{k_3,m-1} & x_{k_3,m} & y_{k_3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{k_{m-1},2} & x_{k_{m-1},3} & \cdots & x_{k_{m-1},m-1} & x_{k_{m-1},m} & y_{k_{m-1}} \end{pmatrix}.$$

Применив алгоритм прямого хода метода Гаусса, преобразуем матрицу $\mathbf{A}_{(k_1,\ldots,k_{m-1})}$ к ступенчатому виду

$$\mathbf{A}'_{(k_1,\ldots,k_{m-1})} = \begin{pmatrix} 1 & x_{k_1,2} & x_{k_1,3} & \cdots & x_{k_1,m-1} & x_{k_1,m} & y_{k_1} \\ 0 & 1 & x'_{k_2,3} & \cdots & x'_{k_2,m-1} & x'_{k_2,m} & y'_{k_2} \\ 0 & 0 & 1 & \cdots & x'_{k_3,m-1} & x'_{k_3,m} & y'_{k_3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & x'_{k_{m-1},m} & y'_{k_{m-1}} \end{pmatrix}.$$

Используя ступенчатую матрицу $\mathbf{A}'_{(k_1,\ldots,k_{m-1})}$, можно значительно сократить вычислительные затраты на нахождение всех узловых точек, лежащих на прямой $l_{(k_1,\ldots,k_{m-1})}$. Действительно, для каждой искомой узловой точки имеем расширенную матрицу

$$\mathbf{A}_{(k_1,\ldots,k_{m-1},i)} = \begin{pmatrix} 1 & x_{k_1,2} & x_{k_1,3} & \cdots & x_{k_1,m-1} & x_{k_1,m} & y_{k_1} \\ 0 & 1 & x'_{k_2,3} & \cdots & x'_{k_2,m-1} & x'_{k_2,m} & y'_{k_2} \\ 0 & 0 & 1 & \cdots & x'_{k_3,m-1} & x'_{k_3,m} & y'_{k_3} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & x'_{k_{m-1},m} & y'_{k_{m-1}} \\ 1 & x_{i,2} & x_{i,3} & \cdots & x_{i,m-1} & x_{i,m} & y_i \end{pmatrix}, \tag{7}$$

где $k_1 < k_2 < \ldots < k_{m-1}$, $i \in \{1, 2, \ldots, n\}$, $i \notin \{k_1, k_2, \ldots, k_{m-1}\}$.

Варьируя номер i в (7), найдем все узловые точки, лежащие на прямой $l_{(k_1,\ldots,k_{m-1})}$.

Спуск с использованием разреженных матриц и с учетом направления спуска. Вычислительную эффективность алгоритма спуска можно повысить, рассматривая направление спуска. Поясним как.

Используя ступенчатую матрицу $\mathbf{A}'_{(k_1,\ldots,k_{m-1})}$ и решив СЛАУ, соответствующую расширенной матрице (7), находим значение m-го коэффициента $a_m^{(k_1,k_2,\ldots k_{m-1},i)}$ ($k_1 < k_2 < \ldots < k_{m-1}$, $i \in \{1,2,\ldots,n\}$, $i \notin \{k_1,k_2,\ldots,k_{m-1}\}$) для каждой узловой точки. После чего по возрастанию $a_m^{(k_1,k_2,\ldots k_{m-1},i)}$ упорядочиваем все узловые точки, которые лежат на прямой $l_{(k_1,\ldots,k_{m-1})}$, и выполняем описанный выше алгоритм спуска, но с учетом направления. Если при непосредственном переходе от одной узловой точки к другой значение целевой функции увеличивается, то в этом направлении значение целевой функции будет увеличиваться во всех узловых точках (вытекает из выпуклости целевой функции). Назовем такое направление «плохим». Для осуществления спуска до вычисления значения целевой функции в очередной узловой точке рассматриваем направление спуска. Если оно «плохое», то переходим к следующей точке, не вычисляя в данной узловой точке ни значение целевой функции, ни значения коэффициентов $a_j^{(k_1,k_2,\ldots k_{m-1},i)}$, $j = 1, 2, \ldots, m-1$.

Анализ вычислительных затрат алгоритмов спуска по узловым прямым

Воспользуемся методом статистических испытаний Монте-Карло [14]. В таблице для 1 000 испытаний приведены средние значения общего количества рассмотренных узловых точек L и узловых прямых P (переходов с одной прямой на другую) в алгоритме спуска с использованием разреженных матриц и с учетом направления спуска для некоторых значений n и m.

Средние количества рассмотренных узловых точек и узловых прямых в алгоритме спуска с использованием разреженных матриц и с учетом направления спуска

n	Кол-во узловых точек L					Кол-во узловых прямых P				
	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$
30	56	91	127	168	211	5,2	6,6	7,7	8,8	9,6
50	85	136	197	262	334	6,1	8,0	9,8	11,3	12,7
100	148	235	348	459	604	7,3	9,9	12,3	14,3	16,8
150	214	337	490	655	847	7,9	10,8	13,7	16,3	19,0
300	397	618	895	1205	1529	9,1	12,7	16,1	19,6	22,9
500	644	1002	1441	1905	2477	9,8	13,9	18,0	21,8	26,0
700	881	1417	2008	2665	3390	10,4	15,0	19,1	23,5	28,1
900	1123	1752	2505	3330	4242	10,9	15,5	20,0	24,4	29,3
1000	1257	1946	2771	3676	4701	11,0	15,7	20,3	25,0	30,1
1200	1461	2341	3345	4400	5626	11,3	16,3	21,0	26,1	31,2
1500	1855	2928	4043	5413	6973	11,8	16,8	21,7	26,9	32,4
1700	2155	3294	4600	6165	7860	12,0	17,1	22,1	27,6	33,1
1850	2278	3520	5004	6763	8602	12,1	17,4	22,5	28,0	33,5
2000	2453	3867	5363	7215	9204	12,2	17,6	22,8	28,2	33,8

Оценим вычислительные затраты спуска с использованием разреженных матриц и с учетом направления спуска. Для этого необходимо оценить средние количества рассмотренных узловых точек L и узловых прямых P. Анализ полученных результатов для различных n и m показал, что $L \sim O(mn)$, $P \sim O(m \ln n)$.

Теорема 3. Алгоритм спуска по узловым прямым имеет вычислительную сложность

$$W = O(m^2 n^2 + m^4 n \ln n + m^2 n \ln^2 n).$$

Доказательство теоремы 3. Оценим вычислительную сложность всех основных функций, заложенных в базисе алгоритма спуска по узловым прямым. Его схема приведена на рис. 3.

Функция trapezoidalMatrix расширенную матрицу размера $(m-1)\times(m+1)$ приводит к трапециевидному виду за $(m-1)$ итерацию. В ходе *i*-й итерации выполняется $(m+2-i)(2m-1-2i)$ операций умножения и вычитания. Следовательно, общее количество выполненных операций

$$I=\sum_{i=1}^{m-1}(m+2-i)(2m-1-2i)=2\sum_{i=1}^{m-1}(m+2-i)^2-5\sum_{i=1}^{m-1}(m+2-i)=\left[\,j=m+2-i\,\right]=2\sum_{j=3}^{m+1}j^2-5\sum_{j=3}^{m+1}j\,.$$

Известно, что $\sum_{j=1}^{m}j^2=\frac{m(m+1)(2m+1)}{6}$ и $\sum_{j=1}^{m}j=\frac{m(m+1)}{2}$, поэтому

$$I=2\cdot\left(\frac{(m+1)(m+2)(2m+3)}{6}-5\right)-5\cdot\left(\frac{(m+1)(m+2)}{2}-3\right)=O(m^3)\,.$$

Функция getMCoeff к трапециевидной матрице размера $(m-1)\times(m+1)$ добавляет строку и вычисляет *m*-й коэффициент. Здесь выполняется $(m-1)$ итераций. В ходе *i*-й итерации выполняется $2(m+1-i)$ операций умножения и вычитания (кроме первой итерации, во время которой выполняется не 2*m*, а *m* операций умножения и вычитания). Следовательно, общее количество выполненных операций будет равно $2\sum_{i=1}^{m-1}(m+1-i)+m-1=m^2-1=O(m^2)$.

Функция obj вычисляет и возвращает значение целевой функции $Q(\mathbf{a})$. Очевидно, что функция obj имеет $O(mn)$ вычислительную сложность.

Функция sort реализует сортировку полученного массива. Известно, что сортировка Хоара в среднем имеет для входного массива из *n* элементов $O(n\ln n)$ вычислительную сложность [15].

Функция descent, имея матрицу вида (7) и *m*-й коэффициент, вычисляет остальные коэффициенты и вызывает функцию obj. Данные действия выполняются циклически, количество итераций примерно равно $L/(m\cdot P)=O(mn)/(m\cdot)(m\ln n))=O(n/(m\ln n))$.

Для вычисления остальных коэффициентов выполняется $(m-1)$ итераций. В ходе *i*-й итерации выполняется 2*i* операций умножения и вычитания. Следовательно, общее количество выполненных операций $2\sum_{i=1}^{m-1}i=m^2-m=O(m^2)$.

Поскольку функция obj имеет вычислительную сложность $O(mn)$, то функция descent имеет $O(n\cdot(mn+m^2)/m\ln n)$ вычислительную сложность.

Теперь из схемы работы алгоритма получим, что спуск по узловым прямым имеет вычислительную сложность

$$O\left\{m\ln n\cdot\left[m\cdot\left(m^3+(n-m)\cdot m^2+(n-m)\cdot\ln(n-m)+\frac{n}{m\ln n}\cdot\left(m^2+mn\right)\right)\right]\right\}=$$
$$=O(m^2n^2+m^4n\ln n+m^2n\ln^2 n+m^3n)=O(m^2n^2+m^4n\ln n+m^2n\ln^2 n)\,.$$

Отметим, что для $n>\max\left(\ln^2 n;\, m^2\ln n\right)$ вычислительная сложность спуска по узловым прямым $W=O(m^2n^2)$.

Выводы

Предложены точные алгоритмы реализации метода наименьших модулей при оценивании параметров линейных регрессионных моделей, основанные на спуске по узловым прямым.

Данные алгоритмы позволяют значительно снизить вычислительные затраты при точной реализации метода наименьших модулей. Выигрыш достигается за счет того, что вместо перебора узловых точек осуществляется спуск по ним. Вычислительная сложность спуска по узловым прямым позволяет на практике реализовать точное оценивание методом наименьших модулей для любых размеров экспериментальных данных.

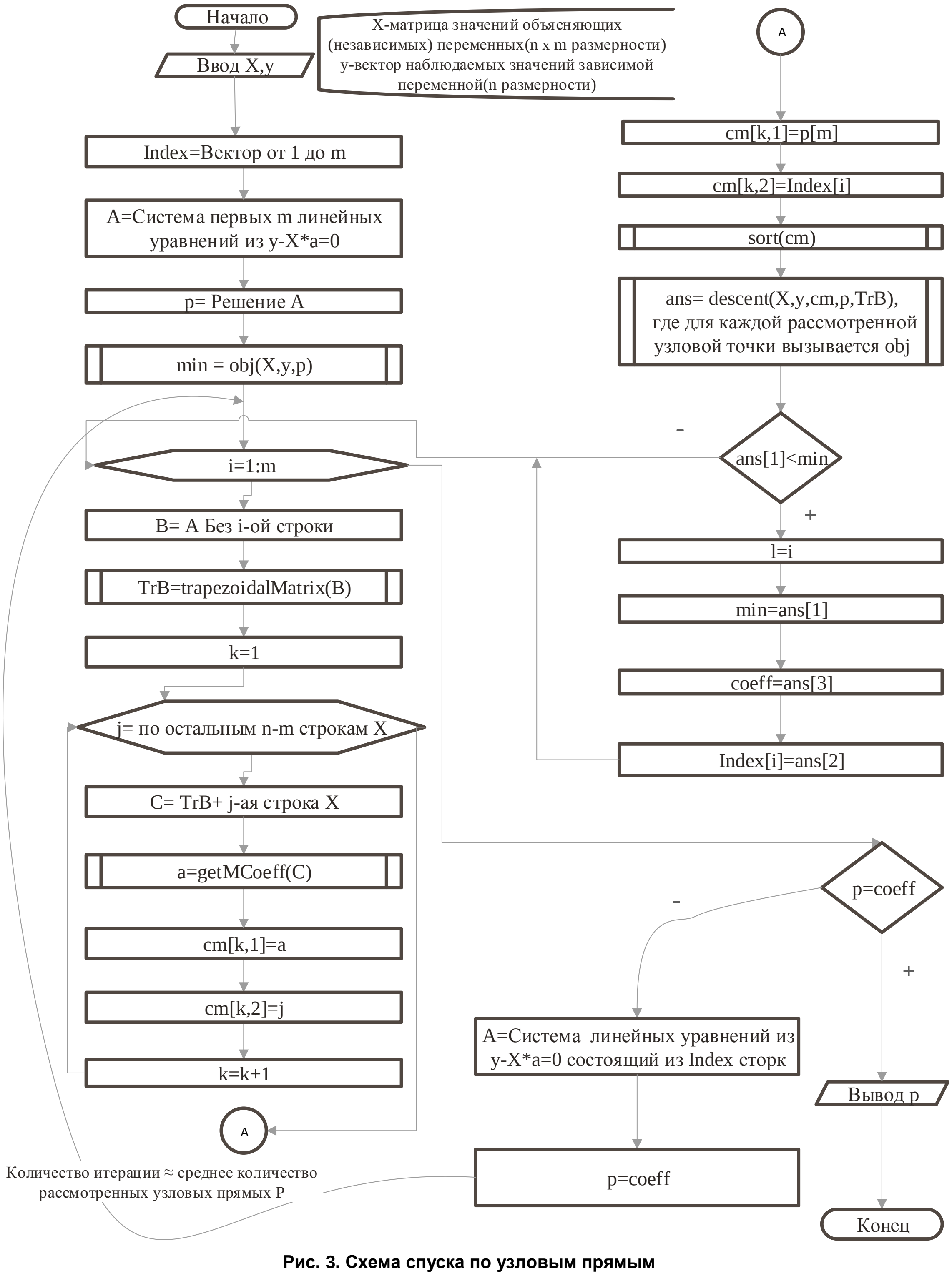

Рис. 3. Схема спуска по узловым прямым

Литература

1. Кендэл, М.Д. Статистические выводы и связи / М.Д. Кендэл, А. Стьюарт. – М.: Наука, 1973. – 899 с.
2. Справочник по теории вероятностей и математической статистике / В.С. Королюк, Н.И. Портенко, А.В. Скороход, А.Ф. Турбин. – М.: Наука, 1985. – 640 с.
3. Демиденко, Е.З. Линейная и нелинейная регрессия / Е.З. Демиденко. – М.: Финансы и статистика, 1981. – 302 с.

4. Айвазян, С.А. Прикладная статистика: Исследование зависимостей / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1985. – 487 с.

5. Bloomfield, P. Least absolute seviations: theory, applications, and algorithms / P. Bloomfield, W.L. Steiger. – Boston-Basel-Stuttgart: Birkhauser, 1983. – 349 p.

6. Мудров, В.И. Методы обработки измерений: квазиправдоподобные оценки / В.И. Мудров, В.Л. Кушко. – М.: Радио и связь, 1983. – 304 с.

7. Weiszfeld, E. On the point for which the sum of the distances to n given points is minimum / E. Weiszfeld // Annals of Operations Research. – 2009. – Vol. 167, Issue 1. – P. 7–41.

8. Акимов, П.А. Уровни неоптимальности алгоритма Вейсфельда в методе наименьших модулей / П.А. Акимов, А.И. Матасов // Автоматика и телемеханика. – 2010. – № 2. – С. 4–16.

9. Тырсин, А.Н. Оценивание линейных регрессионных уравнений с помощью метода наименьших модулей / А.Н. Тырсин, К.Е. Максимов // Заводская лаборатория. Диагностика материалов. – 2012. – Т. 78, № 7. – С. 65–71.

10. Boyd, S. Convex optimization / S. Boyd, L. Vandenberghe. – Cambridge: Cambridge University Press, 2004. – 730 p.

11. Зуховицкий, С.И. Линейное и выпуклое программирование / С.И. Зуховицкий, Л.И. Авдеева. – М.: Наука, 1967. – 460 с.

12. Панюков, А.В. Применение массивно-параллельных вычислений для решения задач линейного программирования с абсолютной точностью / А.В. Панюков, В.В. Горбик // Автоматика и телемеханика. – 2012. – № 2. – С. 73–88.

13. Рокафеллар, Р. Выпуклый анализ / Р. Рокафеллар. – М.: Мир, 1973. – 469 с.

14. Михайлов, Г.А. Численное статистическое моделирование. Методы Монте-Карло / Г.А. Михайлов, А.В. Войтишек. – М.: Академия, 2006. – 366 с.

15. Алгоритмы: построение и анализ / Т.Х. Кормен, Ч. Лейзерсон, Р. Ривест, К. Штайн. – М.: Вильямс, 2013. – 1323 с.

Поступила в редакцию 16 января 2017 г.

Bulletin of the South Ural State University
Series “Mathematics. Mechanics. Physics”
2018, vol. 10, no. 2, pp. 47–56

DOI: 10.14529/mmph180205

EXACT EVALUATION OF LINEAR REGRESSION MODELS BY THE LEAST ABSOLUTE DEVIATIONS METHOD BASED ON THE DESCENT THROUGH THE NODAL STRAIGHT LINES

A.N. Tyrsin[1,2], A.A. Azaryan[2]
[1]Science and Engineering Center «Reliability and Resource of Large Systems and Machines», Ural Branch, Russian Academy of Sciences, Yekaterinburg, Russian Federation
[2]Ural Federal University named after the first President of Russia B.N.Yeltsin, Yekaterinburg, Russian Federation
E-mail: at2001@yandex.ru

When building linear models, in many cases one has to deal with stochastic nonhomogeneity of experimental data. This manifests itself in violation of the assumptions of the Gauss–Markov theorem, in particular, observations can contain outliers. Under these circumstances the estimation of the parameters of models is required to be performed using resistant methods. Among those is the least absolute deviations method. However, the known algorithms for its implementation are sufficiently effective only for small dimensions of models and a limited volume of samples. The purpose of this study is the development of effective computational algorithms for implementation of the least absolute deviations method, which have no limitations as to the order of models, and the amount of experimental data. Algorithms for the exact solution of the problem on estimating the parameters of linear regression models by the least absolute deviations method are described. They are based on the descent through the nodal straight lines. To reduce computational costs, the particular feature of nodal straight lines is used – all nodes located on each such straight line are intersections of a set of hyperplanes, of which only one hyperplane is different. These algorithms significantly outperform the best-known brute-force search and can be effectively used in practice. The computational complexity of the descent algorithm for nodal straight lines is assessed. The scheme of the algorithm is provided.

Keywords: the least absolute deviations method; linear regression model; algorithm; nodal point; nodal straight line; hyperplane; computational complexity.

References

1. Kendel M.D., St'yuart A. *Statisticheskie vyvody i svyazi* (Statistical Inferences and Relationships). Moscow, Nauka Publ., 1973, 899 p. (in Russ.).

2. Korolyuk V.S., Portenko N.I., Skorokhod A.V., Turbin A.F. *Spravochnik po teorii veroyatnostey i matematicheskoy statistike* (A Handbook on Probability Theory and Mathematical Statistics). Moscow, Nauka Publ., 1985, 640 p. (in Russ.).

3. Demidenko E.Z. *Lineynaya i nelineynaya regressiya* (Linear and nonlinear regression). Moscow, Finansy i statistika Publ., 1981, 302 p. (in Russ.).

4. Ayvazyan S.A., Enyukov I.S., Meshalkin L.D. *Prikladnaya statistika: Issledovanie zavisimostey* (Applied statistics: Dependency Studies). Moscow, Finansy i statistika Publ., 1985, 487 p. (in Russ.).

5. Bloomfield P., Steiger W.L. *Least absolute seviations: theory, applications, and algorithms*. Boston-Basel-Stuttgart, Birkhauser, 1983, 349 p.

6. Mudrov V.I., Kushko V.L. *Metody obrabotki izmereniy: kvazipravdopodobnye otsenki* (Measurement processing methods: quasi-truth estimates). Moscow, Radio i svyaz' Publ., 1983, 304 p. (in Russ.).

7. Weiszfeld E. On the point for which the sum of the distances to n given points is minimum. *Annals of Operations Research*, 2009, Vol. 167, no. 1, pp. 7–41. DOI: 10.1007/s10479-008-0352-z

8. Akimov P.A., Matasov A.I. Levels of nonoptimality of the Weiszfeld Algorithm in the least-modules method. *Automation and Remote Control*, 2010, Vol. 71, Issue 2, pp. 172–184. DOI: 10.1134/S0005117910020025

9. Tyrsin A.N., Maksimov K.E. Estimation of the Linear Regression Equations Using the Least-Modules Method. *Zavodskaya laboratoriya. Diagnostika materialov*, 2012, Vol. 78, no. 7, pp. 65–71. (in Russ.).

10. Boyd S., Vandenberghe L. *Convex optimization.* Cambridge, Cambridge University Press, 2004, 730 p. DOI: 10.1017/CBO9780511804441

11. Zukhovitskiy S.I., Avdeeva L.I. *Lineynoe i vypukloe programmirovanie* (Linear and convex programming). Moscow, Nauka Publ., 1967, 460 p.

12. Panyukov A.V., Gorbik V.V. *Using massively parallel computations for absolutely precise solution of the linear programming problems*. *Automation and Remote Control*, 2012, Vol. 73, Issue 2, pp. 276–290. DOI: 10.1134/S0005117912020063

13. Rokafellar R. *Vypuklyy analiz* (Convex analysis). Moscow, Mir Publ., 1973, 469 p. (in Russ.).

14. Mikhaylov G.A., Voytishek A.V. *Chislennoe statisticheskoe modelirovanie. Metody Monte-Karlo* (Numerical statistical modeling. Monte Carlo methods). Moscow, Akademiya Publ., 2006, 368 p. (in Russ.).

15. Kormen T.Kh., Leyzerson Ch., Rivest R., Shtayn K. *Algoritmy: postroenie i analiz* (Algorithms: construction and analysis). Moscow, Vil'yams Publ., 2013, 1323 p. (in Russ.).

Received January 16, 2017