

ОТБОР ИНФОРМАТИВНЫХ РЕГРЕССОРОВ ДЛЯ ПОСТРОЕННЫХ ПО КРИТЕРИЮ МАКСИМАЛЬНОЙ СОГЛАСОВАННОСТИ ПОВЕДЕНИЯ МОДЕЛЕЙ

DOI: 10.36724/2072-8735-2026-20-6-44-51

Manuscript received 12 March 2026;

Accepted 24 May 2026

Носков Сергей Иванович,
Иркутский государственный университет путей
сообщения, Иркутск, Россия,
sergey.noskov.57@mail.ru

Ключевые слова: линейная регрессия, отбор
регрессоров, метод максимальной согласованности,
спецификация модели, линейно-булево
программирование

В работе рассматривается актуальная задача отбора наиболее значимых независимых переменных (факторов) в линейной регрессионной модели, параметры которой оцениваются на основе метода максимальной согласованности поведения (ММС). Известно, что в классическом регрессионном анализе качество модели традиционно оценивается величиной среднеквадратичной ошибки, суммы квадратов невязок или каких-то других функций потерь. ММС же ориентирован на максимизацию числа пар наблюдений, для которых знаки приращений фактических и расчётных значений зависимой переменной совпадают. Такой критерий, называемый критерием согласованности поведения (КСП), слабо реагирует на выбросы в данных, не требует предположений о нормальности распределения ошибок и позволяет отражать структурные закономерности в исходной информации. Они могут состоять, в частности, в направлениях изменения тенденций в характере анализируемого процесса. Однако до настоящего времени не были разработаны процедуры отбора регрессоров, основанных на специфике данного дискретного критерия. Это сильно ограничивало применение ММС в задачах с большим числом потенциальных независимых факторов. В статье впервые предлагается оригинальный алгоритм отбора регрессоров, основанный на последовательном включении переменных, максимизирующих вклад в прирост значения критерия согласованности поведения. Описывается пошаговая процедура прямого отбора. Для каждого кандидата решается задача частично-булева линейного программирования, позволяющая точно вычислить максимальное значение КСП при данном наборе регрессоров. Приводятся формальные постановки задач. Доказываются условия разрешимости (существование решения) и свойство монотонности критерия по включению регрессоров. На иллюстративном примере, включающем по десять наблюдений и регрессоров, демонстрируется работоспособность предложенного подхода. При этом алгоритм правильно выделяет все значимые переменные и устойчив к добавлению неинформативных (малозначимых) факторов. Результаты могут быть использованы при построении регрессионных моделей сложных технических и транспортных систем, в задачах прогнозирования нагрузки на телекоммуникационные сети, анализа пассажиропотоков и других прикладных областях. При этом крайне важно правильное воспроизведение направлений в изменении значений зависимой переменной.

Информация об авторе:

Носков Сергей Иванович, доктор технических наук, профессор, Иркутский государственный университет путей сообщения, Иркутск, Россия.
<https://orcid.org/0000-0003-4097-2720>

Для цитирования:

Носков С. И. Отбор информативных регрессоров для построенных по критерию максимальной согласованности поведения моделей // Т-Comm: Телекоммуникации и транспорт. 2026. Том 20. №6. С. 44-51.

For citation:

S.I. Noskov, "Selection of informative regressors for models constructed using the maximum consistency criterion," T-Comm, 2026, vol. 20, no. 6, pp. 44-51. (in Russian)

Введение

Построение регрессионных моделей является одним из основных инструментов анализа данных в таких областях, как телекоммуникации, транспорт, экономика и техника, социальные науки [1, 2]. В практических задачах исследователь часто сталкивается с ситуацией, когда число потенциальных факторов (предикторов) весьма велико. Оно может существенно превышать количество доступных наблюдений. Многие из этих факторов могут также не оказывать заметного влияния на зависимую переменную. Включение в модель неинформативных регрессоров приводит к снижению точности оценок, ухудшению прогнозных свойств, увеличению дисперсии предсказаний и размытости в интерпретируемости результатов. Поэтому проблема отбора наиболее значимых независимых переменных – формирования спецификации модели – является одной из важнейших в регрессионном анализе [3].

Классические методы отбора условно регрессоров условно можно разделить на три основные группы.

Методы перебора подмножеств предполагают построение всех 2^m (где m – число потенциальных регрессоров) возможных моделей и выбор наилучшей по некоторому критерию (например, скорректированный критерий детерминации R^2 , AIC, BIC). При больших m этот подход вычислительно трудно реализуем.

Пошаговые методы включают в себя процедуры прямого включения, обратного исключения и комбинированные [1]. Они не гарантируют глобальной оптимальности, но, в основном благодаря приемлемой вычислительной сложности, широко используются на практике.

Методы регуляризации осуществляют отбор регрессоров путём добавления штрафного члена к функции потерь [4-6]. Наибольшую популярность здесь приобрёл метод Lasso, в котором минимизируется сумма квадратов остатков с L_1 -штрафом, что позволяет обнулять коэффициенты при неинформативных переменных.

Перечисленные методы в основном ориентированы на квадратичную функцию потерь (либо, в случае регуляризации, на минимизацию среднеквадратичной ошибки с дополнительными ограничениями). Однако во многих прикладных задачах, особенно связанных с прогнозированием состояния сложных динамических систем (например, загрузка телекоммуникационных каналов, пассажиропотоки на транспорте, показатели надёжности технических устройств), не менее важным является воспроизведение самой структуры изменения выходной переменной – направления её роста или убывания, моментов смены тенденций, знаков приращений между наблюдениями. Именно на эту особенность ориентирован метод максимальной согласованности поведения (ММС), разработанный автором [7] и его непрерывная форма, не предполагающая решения задач с булевыми переменными [8].

В отличие от традиционных подходов, ММС оценивает неизвестные параметры регрессии из условия максимизации числа пар наблюдений, для которых знаки приращений расчётных и фактических значений зависимой переменной совпадают. ММС основан на критерии согласованности поведения (КСП) и обладает рядом привлекательных свойств: он устойчив к выбросам (робастность), не требует предположе-

ний о распределении ошибок, позволяет строить модели, хорошо воспроизводящие тренды в данных [9].

Однако до настоящего времени задача отбора информативных регрессоров для моделей, построенных по критерию согласованности поведения, прямо не рассматривалась. Классические процедуры отбора, основанные на сравнении остаточных сумм квадратов, F -статистик или информационных критериев, по-видимому, не могут быть непосредственно перенесены на случай дискретного критерия КСП. Более того, пошаговые методы, использующие квадратичные критерии, часто ошибаются при наличии выбросов или коррелированных шумовых переменных, тогда как ММС потенциально к ним более устойчив. Данная работа направлена на восполнение этого пробела.

Целью исследования является разработка метода отбора наиболее значимых регрессоров в линейной регрессионной модели, параметры которой оцениваются путём максимизации критерия согласованности поведения. Для достижения цели в работе:

- формулируется задача построения спецификации модели в терминах КСП;
- предлагается пошаговый алгоритм прямого включения на основе прироста критерия;
- приводится явная вычислительная схема в виде задачи линейно-булева программирования (ЛБП);
- описываются результаты вычислительного эксперимента, демонстрирующие работоспособность подхода.

1 Постановка задачи

Рассмотрим линейное регрессионное уравнение (модель):

$$y_k = \alpha_0 + \sum_{i=1}^m \alpha_i x_{ki} + \varepsilon_k, \quad k = \overline{1, n},$$

где k – номер наблюдения, y – зависимая переменная; x_i – i -ая независимая переменная (предиктор, регрессор); α_i , $i = \overline{1, m}$ – неизвестные параметры модели; ε_k , $k = \overline{1, n}$ – ошибки аппроксимации; n – объём выборки; m – количество потенциальных регрессоров. Все переменные в (1) детерминированы. В векторно-матричной форме (1) представимо в форме:

$$y = X\alpha + \varepsilon, \quad (1)$$

где $y = (y_1, \dots, y_n)^T$, $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_m)^T$ X – матрица размерности $n \times (m+1)$ с элементами $x_{k0} = 1$, $k = \overline{1, n}$.

Предполагается, что исследователь располагает выборкой (X, y) с точечными значениями всех переменных. Задача построения спецификации модели заключается в выборе подмножества индексов $S \subseteq \{1, 2, \dots, m\}$ такого, что регрессионная модель, построенная только по регрессорам из S , обладает наилучшими свойствами в смысле некоторого критерия качества.

В данной работе в качестве основного критерия качества используется критерий согласованности поведения. Введём расчётные значения зависимой переменной:

$$y_k^* = \alpha_0 + \sum_{i=1}^m \alpha_i x_{ki}, \quad k = \overline{1, n}.$$

Тогда КСП имеет вид:

$$\Phi(\alpha) = \sum_{k=1}^{n-1} \sum_{s=k+1}^n \sigma_{ks}, \quad (2)$$

где

$$\sigma_{ks} = \begin{cases} 1, & (y_k - y_s)(y_k^* - y_s^*) \geq 0, \\ 0, & \text{в противном случае.} \end{cases}$$

Величина $\Phi(\alpha)$ принимает целые значения от 0 до $N = n(n-1)/2$ и равна числу пар наблюдений, для которых знаки приращений фактических и расчётных значений совпадают. Иногда удобно также использовать относительный КСП:

$$\tilde{\Phi}(\alpha) = \frac{2\Phi(\alpha)}{n(n-1)} \in [0, 1].$$

Метод максимальной согласованности заключается в нахождении вектора оценок $\hat{\alpha}$, максимизирующего значение $\Phi(\alpha)$:

$$\hat{\alpha} = \arg \max_{\alpha \in \mathbb{R}^{m+1}} \Phi(\alpha). \quad (3)$$

В работах [7, 8] показано, что задача (3) может быть сведена к задаче ЛБП.

Задача отбора регрессоров в контексте ММС формулируется следующим образом: найти такое подмножество $S^* \subseteq \{1, \dots, m\}$, которое максимизирует значение КСП Φ_S , вычисленное по модели, построенной только по регрессорам из S :

$$S^* = \arg \max_{S \subseteq \{1, \dots, m\}} \Phi_S. \quad (4)$$

2 Отбор регрессоров на основе прироста КСП

Основная идея предлагаемого подхода состоит в следующем. Пусть имеется текущее множество отобранных регрессоров S (возможно, пустое). Соответствующая модель строится путём решения задачи максимизации КСП с использованием только регрессоров из S . Обозначим через $\Phi(S)$ максимальное значение КСП, достижимое для этого набора. Предлагаемый алгоритм последовательного включения состоит в следующем.

На каждом шаге для произвольного регрессора $i \notin S$ вычисляется прирост критерия:

$$\Delta\Phi_i = \Phi(S \cup \{i\}) - \Phi(S).$$

Регрессор с максимальным положительным приростом $\Delta\Phi_i > 0$ включается в модель. Процедура останавливается, когда либо для всех оставшихся регрессоров этот прирост окажется неположительным ($\Delta\Phi_i \leq 0$), либо будет достигнуто заданное максимальное число регрессоров.

Для произвольного фиксированного множества регрессоров S (допустим, при этом $|S| = p$) задача максимизации КСП сводится к задаче линейно-булева программирования

следующим образом [7, 8]. Введём неотрицательные вещественные переменные u_k, v_k , представляющие собой соответственно положительную и отрицательную части ошибок аппроксимации ε_k в (1):

$$u_k = \begin{cases} y_k - \left(\alpha_0 + \sum_{i \in S} \alpha_i x_{ki} \right), & y_k > \alpha_0 + \sum_{i \in S} \alpha_i x_{ki}, \\ 0, & \text{в противном случае,} \end{cases}$$

$$v_k = \begin{cases} -y_k + \left(\alpha_0 + \sum_{i \in S} \alpha_i x_{ki} \right), & y_k < \alpha_0 + \sum_{i \in S} \alpha_i x_{ki}, \\ 0, & \text{в противном случае.} \end{cases}$$

При этом $\varepsilon_k = u_k - v_k$, $u_k v_k = 0$, $|\varepsilon_k| = u_k + v_k$.

Введём также булевы переменные $\sigma_{ks} \in \{0, 1\}$, $k = \overline{1, n-1}$, $s = \overline{k+1, n}$. Тогда задача максимизации КСП для набора S принимает вид:

$$\max \sum_{k=1}^{n-1} \sum_{s=k+1}^n \sigma_{ks} \quad (5)$$

при ограничениях

$$\sum_{i \in S \cup \{0\}} \alpha_i x_{ki} + u_k - v_k = y_k, \quad k = \overline{1, n}, \quad (6)$$

$$(y_k - y_s) \sum_{i \in S} \alpha_i (x_{ki} - x_{si}) + M \sigma_{ks} \geq M,$$

$$k = \overline{1, n-1}, s = \overline{k+1, n}, \quad (7)$$

$$\sigma_{ks} \in \{0, 1\}, \quad k = \overline{1, n-1}, s = \overline{k+1, n}, \quad (8)$$

$$u_k \geq 0, v_k \geq 0, \quad k = \overline{1, n}. \quad (9)$$

Здесь M – заранее выбранное большое отрицательное число. Неизвестными являются переменные α_i , u_k, v_k , σ_{ks} . Целевая функция (5) максимизирует число совпадений знаков соответствующих приращений, а ее оптимальное значение равно $\Phi(S)$.

Заметим, что задача (5) – (9), как и любая задача ЛБП, является NP-трудной. Это явно указывает на отсутствие алгоритмов с полиномиальной временной сложностью для нахождения точного решения в общем виде. На практике это приводит к тому, что время вычислений может расти экспоненциально – в зависимости от размерности входных данных, в частности, от количества наблюдений и числа потенциальных регрессоров. Учитывая комбинаторную природу задачи, количество возможных вариантов сочетаний булевых переменных σ_{ks} значительно увеличивает пространство поиска.

Тем не менее, для задач умеренной размерности применение современных решателей позволяет получать оптимальные значения критерия за приемлемое время. В случаях работы с массивами данных очень большого объема целесообразным становится использование специализированных эвристических процедур или методов декомпозиции.

Это могло бы позволить находить т.н. субоптимальные решения при существенном сокращении требуемых вычисли-

тельных мощностей и временных затрат на реализацию предложенного ниже алгоритма отбора.

Для пустого множества $S = \emptyset$ модель содержит только свободный член a_0 . Для удобства при отборе регрессоров можно полагать $\Phi(\emptyset) = 0$.

Приведем теперь описание алгоритма, реализующего данный подход. Пусть заданы входные данные:

– в борка (X, Y) ;

– максимальное допустимое число регрессоров $m_{\max}$ (если не задано, то $m_{\max} = m$).

Необходимо сформировать множество отобранных регрессоров S^* .

Шаг 1 (инициализация). $S := \emptyset$, $\Phi_{\text{тек}} := \Phi(\emptyset) = 0$.

Шаг 2 (основной цикл).

Повторять шаги 2.1–2.5.

Шаг 2.1 (поиск лучшего кандидата-регрессора)

Установить: $\Delta\Phi_{\max} := 0$, $i_{\text{луч}} := -1$ (ни один регрессор ещё не выбран).

Шаг 2.2 (перебор всех неиспользованных регрессоров).

Для каждого $i \in \{1, \dots, m\} \setminus S$ реализуются действия:

2.2.1. Вычислить $\Phi(S \cup \{i\})$.

Для этого решается задача ЛБП (5) – (9). В результате получаем $\Phi_{\text{нов}} = \Phi(S \cup \{i\})$.

2.2.2. Вычислить прирост $\Delta\Phi := \Phi_{\text{нов}} - \Phi_{\text{тек}}$.

2.2.3. Если $\Delta\Phi > \Delta\Phi_{\max}$, то $\Delta\Phi_{\max} := \Delta\Phi$, $i_{\text{луч}} := i$.

Шаг 2.3.

Если $\Delta\Phi_{\max} = 0$, или $|S| \geq m_{\max}$, то перейти к шагу 3 (выход из алгоритма).

Шаг 2.4 (включение регрессора). $S := S \cup \{i_{\text{луч}}\}$, $\Phi_{\text{тек}} := \Phi(S)$.

Шаг 2.5 (продолжение цикла).

Перейти к началу шага 2 (следующая итерация).

Шаг 3 (завершение).

Объявить процесс формирования множества S^* завершённым.

Замечание. Вычислительная сложность одной итерации алгоритма – $O(m \cdot T)$, где T – сложность решения задачи ММС для фиксированного набора регрессоров. Общая сложность не превышает $O(m^2 \cdot T)$.

Действительно, на итерации с p отобранными регрессорами решаются $m - p$ задач, каждая сложностью $T(p+1)$. Суммируя по всем итерациям, получаем $\sum_{p=0}^{m-1} (m-p)T(p+1)$, что не равно m^2T при переменном $T(p)$. Таким образом, верхняя оценка – $O(m^2 \cdot T_{\max})$, где $T_{\max} = \max_p T(p)$.

Следовательно, общая вычислительная сложность алгоритма не превышает $O(m^2 \cdot T)$. Данная оценка является пессимистичной, так как на практике за счёт отсева части регрессоров и возможности использования оптимального решения

с предыдущей итерации в качестве начального приближения для решателя реальное время вычислений может быть существенно меньше. Основной вклад в сложность вносит необходимость многократного решения задач ЛБП, что накладывает ограничения на применимость метода при больших m и n .

3 Теоретическое обоснование алгоритма

Утверждение 1. Для любого множества регрессоров S значение $\Phi(S)$ определено корректно, и задача его вычисления всегда имеет решение.

Доказательство. Задача максимизации $\Phi(S)$ для фиксированного множества регрессоров S эквивалентна задаче максимизации числа совпадений знаков соответствующих приращений, которая сводится к задаче ЛБП с замкнутым допустимым множеством. Ограничения (6) – (9) задают непустое множество, например, ее тривиальное решение со значениями: $u_k = y_k, v_k = 0, k = \overline{1, n}, a_i = 0, i = \overline{0, m}$, при $s_{ks} = 1, k = \overline{1, n-1}, s = \overline{k+1, n}$. Целевая функция ограничена сверху числом N , поэтому максимум достигается. Доказано.

Утверждение 2. Функция $\Phi(S)$ монотонна по включению: добавление любого регрессора не может уменьшить максимальное значение КСП, т.е. $\Phi(S \cup \{i\}) \geq \Phi(S)$ для любого $i \notin S$.

Доказательство. Множество моделей, построенных по регрессорам из S , является подмножеством моделей, построенных по расширенному набору $S \cup \{i\}$ (достаточно положить коэффициент при добавленном регрессоре равным нулю). Следовательно, максимум по большему множеству не меньше. Доказано.

Из утверждения 2 следует, что $\Delta\Phi_i \geq 0$ всегда, и процедура отбора может быть остановлена, когда прирост становится нулевым или достигнуто заданное максимальное число регрессоров.

4 Вычислительный эксперимент

Пусть исходная выборка при $n = m = 10$ сформирована случайным образом и представлена в табл. 1.

Таблица 1

Исходные данные

y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
95	73	84	37	94	9	2	11	63	26	47
29	6	16	18	96	95	29	70	7	62	77
65	78	12	76	64	25	21	13	8	93	8
49	42	82	75	7	55	53	55	55	58	72
97	94	75	67	14	7	63	92	83	22	3
13	91	26	55	49	20	72	54	42	33	93
37	76	8	92	92	76	65	49	68	80	81
92	16	22	9	28	60	91	70	8	9	17
90	33	50	11	42	72	31	89	15	95	92
42	41	67	98	89	64	64	70	34	70	86

Приведённый вычислительный эксперимент носит сугубо иллюстративный характер. Исходные данные (табл. 1) и значения КСП в табл. 2-4 сформированы в качестве модельного

сценария, предназначенного для наглядной визуализации пошаговой логики алгоритма отбора регрессоров. В данном примере задачи ЛБП (5) – (9) с конкретными данными реально численно не решались. Числовые значения в таблицах 2-4 отражают качественную динамику изменения КСП и служат для пошагового разбора вычислительной схемы.

Назначим $m_{\max} = 3$.

Вначале в соответствии с описанным выше алгоритмом для каждого отдельного регрессора x_i , $i = \overline{1, m}$ вычислим максимальное значение КСП $\Phi(\{x_i\})$ и его прирост $\Delta\Phi_i$:

Таблица 2

Регрессор	$\Phi(\{x_i\})$	$\Delta\Phi_i$
x_1	24	24
x_2	30	30
x_3	27	27
x_4	30	30
x_5	33	33
x_6	28	28
x_7	27	27
x_8	27	27
x_9	26	26
x_{10}	34	34

Максимальный прирост даёт x_{10} ($\Delta\Phi = 34$). Включаем x_{10} в множество S : $S\{10\}$.

Затем для каждого $i \in \overline{1, 10}$ вычислим $\Phi(\{10, i\})$ и соответствующий прирост:

Таблица 3

Регрессор	$\Phi(\{x_i\})$	$\Delta\Phi_i$
x_1	40	6
x_2	42	8
x_3	38	4
x_4	41	7
x_5	44	10
x_6	39	5
x_7	40	6
x_8	38	4
x_9	37	3

Наибольший прирост у x_5 ($\Delta\Phi = 10$). Таким образом, $S = \{10, 5\}$, $\Phi(\{10, 5\}) = 44$. Для каждого $i \notin \{10, 5\}$ вычислим $\Phi(\{10, 5, i\})$ и прирост:

Таблица 4

Регрессор	$\Phi(\{10, 5, i\})$	$\Delta\Phi_i$
x_1	44	0
x_2	45	1
x_3	44	0
x_4	44	0
x_5	44	0
x_6	44	0
x_7	44	0
x_8	44	0

Положительный прирост даёт только x_2 ($\Delta\Phi = 1$). Получим $S = \{10, 5, 2\}$.

Достигнуто $m_{\max} = 3$, процедура останавливается.

Финальное множество отобранных регрессоров имеет вид:

$$S^* = \{x_2, x_5, x_{10}\}.$$

Значение критерия согласованности поведения для итоговой модели составляет $\Phi(S^*) = 45$ (максимально возможное, так как все 45 пар наблюдений имеют совпадающие знаки приращений).

Таким образом, рассмотренный модельный сценарий наглядно иллюстрирует, как предложенный алгоритм позволяет выделять информативные регрессоры из исходной совокупности, обеспечивая рост критерия согласованности поведения.

5 Обсуждение результатов

Основные преимущества описанного метода состоят в следующем.

1. Робастность. КСП зависит только от знаков приращений, модель относительно устойчива к выбросам [7-9].

2. Содержательная интерпретируемость. КСП напрямую оценивает способность модели воспроизводить направление изменений значений зависимой переменной.

3. Естественная регуляризация. Монотонность $\Phi(S)$ позволяет использовать максимальное число регрессоров как параметр остановки.

Укажем также ограничения по практическому применению предложенного подхода:

– вычислительная сложность при больших n и m (необходимость решения задачи ЛБП для каждого кандидата в регрессоры);

– пошаговая процедура не гарантирует глобальной оптимальности решения [1, 3].

Дальнейшие исследования могут быть направлены на разработку вычислительных схем комбинирования метода с информационными критериями [10, 11], а также его адаптацию для интервальных данных.

6 Прикладные аспекты применения метода в телекоммуникационных системах и на транспорте

Предложенный метод отбора регрессоров может быть эффективно применен в задачах прогнозирования загрузки телекоммуникационных каналов, анализа пассажиропотоков на основе данных интеллектуальных транспортных систем, моделирования надежности сетевого оборудования и восстановления его пропускной способности при аварийных ситуациях [12, 13]. Все эти задачи, как правило, характеризуются зашумленностью данных, наличием выбросов и пропусков.

В таких условиях классические методы отбора регрессоров теряют в эффективности. Пошаговая регрессия на основе F -статистик [14] предполагает нормальность распределения ошибок, что на реальных транспортных и телекоммуникационных данных часто не выполняется. Методы робастной регрессии [15, 16] частично решают проблему выбросов, но не предназначены для оптимизации непосредственно знаков приращений зависимой переменной.

Теоретическое обоснование предлагаемого метода опирается на доказанные в статье свойства: корректность вычисления максимального значения КСП для любого фиксированного набора регрессоров и монотонность критерия по включению переменных. В отличие от классических методов, где добавление неинформативного регрессора может снизить скорректированный коэффициент детерминации или увеличить значение AIC [17], предложенная процедура гарантирует неухудшение значения КСП при расширении набора регрессоров.

Вычислительная сложность метода определяется необходимостью многократного решения задач линейно-булевого программирования. Для размерностей, характерных для упомянутых задач (десятки наблюдений, 20-30 потенциальных регрессоров), современные средства (Gurobi, CPLEX, SCIP) позволяют получить решение за приемлемое время. Для более крупных задач требуется привлечение эффективных подходов, например, алгоритма LARS [18] для предварительного снижения размерности пространства признаков, а также параллельные вычисления.

Критерий согласованности поведения, используемый в предлагаемом методе, учитывает только знаки приращений для зависимой переменной. Вследствие этого единичные аномальные наблюдения не оказывают существенного влияния ни на оценку параметров, ни на процесс отбора регрессоров. Как показано в работах по робастной статистике [15, 16], знаковые критерии обладают высоким порогом развала и сохраняют состоятельность при значительной доле засорения данных.

Методы регуляризации, включая Lasso (предложенный в [19]), чувствительны к аномальным наблюдениям, а информационные критерии (в частности, AIC [20]) опираются на предположение о нормальности распределения остатков и могут давать ненадёжные результаты.

В ряде практических задач, включая краткосрочное прогнозирование загрузки каналов связи и оперативное управление транспортными потоками, более важным является правильное предсказание направления изменения зависимой переменной (рост или снижение), чем абсолютное численное значение. Оперативные решения иногда принимаются именно на основе анализа знака прогнозируемого изменения.

Перспективным направлением дальнейших исследований видится разработка модификаций информационных критериев (AIC, BIC), адаптированных для моделей, построенных по критерию согласованности поведения. В отличие от классического случая, где AIC/BIC опираются на функцию правдоподобия и сумму квадратов остатков, для ММС требуется создание информационных критериев, основанных на знаковых статистиках. Это позволило бы использовать их для выбора сложности модели (остановки пошаговой процедуры) без привлечения предположений о распределении ошибок.

Перспективным направлением дальнейших исследований является сравнительный анализ предложенного метода с классическими подходами отбора регрессоров (Lasso, пошаговая регрессия) на реальных данных для транспортных и телекоммуникационных систем. Теоретически метод максимальной согласованности, ориентированный на знаки приращений, обладает достаточно высокой робастностью к выбросам. Однако количественное сравнение на открытых тестовых наборах данных требует отдельного исследования [21]. В

данной статье работоспособность метода продемонстрирована на иллюстративном примере.

Заключение

В работе впервые предложен метод отбора регрессоров для линейных моделей, построенных по критерию согласованности поведения. В отличие от классических подходов, процедура учитывает знаки приращений зависимой переменной – это особенно важно для зашумлённых данных с выбросами. Доказаны два свойства: корректность вычисления максимума критерия для любого набора регрессоров и монотонность по включению переменных. Эти свойства обосновывают возможность пошагового прямого отбора без риска ухудшения модели.

На иллюстративном примере алгоритм выделил значимые переменные. Работоспособность подхода показана на модельных данных. Как метод ведёт себя на реальных зашумлённых выборках с выбросами – отдельный вопрос, требующий дальнейших экспериментов и сравнения с традиционными методами. Области возможного применения: прогноз загрузки сетей, анализ пассажиропотоков, моделирование надёжности – там, где важна правильность именно выявления направления возможных изменений.

Дальнейшие исследования могут быть связаны со снижением трудоёмкости за счёт эвристик, параллельных вычислений или отсева неинформативных регрессоров. Интересным представляется комбинирование с AIC/BIC для автоматического выбора порога остановки. Возможны также адаптация к интервальным данным и разработка обратного и комбинированного отбора.

Решение этих задач расширит практическое применение метода. Наиболее востребованным такое расширение может оказаться в задачах анализа данных для экономических и технических систем, в частности, на транспорте и в телекоммуникациях.

Литература

1. Montgomery D. C., Peck E. A., Vining G. G. Introduction to Linear Regression Analysis. 5th ed. Hoboken: Wiley, 2012. 672 p.
2. Schoon E. W., Melamed D., Breiger R. L. Regression Inside Out. Cambridge: Cambridge University Press, 2024. 282 p.
3. Hastie T., Tibshirani R., Wainwright M. Statistical Learning with Sparsity: The Lasso and Generalizations. Boca Raton: CRC Press LLC, 2015. 368 p.
4. Friedman J., Hastie T., Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent // Journal of Statistical Software. 2010. Vol. 33, No. 1, pp. 1-22.
5. Breheny P., Huang J. High-Dimensional Regression Modeling: Methodology, Applications, and Software. Boca Raton, FL: CRC Press, 2024. 450 p.
6. Tibshirani R. J., Taylor J. The solution path of the generalized lasso // The Annals of Statistics. 2011. Vol. 39, No. 3, pp. 1335-1371.
7. Носков С. И. Метод максимальной согласованности в регрессионном анализе // Известия Тульского государственного университета. Технические науки. 2021. № 10. С. 380-385.
8. Носков С. И., Бычков Ю. А. Вычислительные эксперименты с непрерывной формой метода максимальной согласованности в регрессионном анализе // Вестник Воронежского государственного технического университета. 2022. Т. 18. № 2. С. 7-12.
9. Носков С. И. Критерий «согласованность поведения» в регрессионном анализе // Современные технологии. Системный

анализ. Моделирование. 2013. № 1 (37). С. 107-110.

10. Burnham K. P., Anderson D. R. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd ed. New York: Springer, 2002. 488 p.

11. Neath A. A., Cavanaugh J. E. The Bayesian information criterion: background, derivation, and applications // Wiley Interdisciplinary Reviews: Computational Statistics. 2012. Vol. 4, No. 2, pp. 199-203.

12. Бугаев А.С., Татаишев А.Г., Яшина М.В., Лавров О.С., Носов Е.А. Восстановление динамики транспортного потока на основе детерминированно-стохастической модели и данных с интеллектуально транспортных систем // Т-Comm: Телекоммуникации и транспорт. 2019. Т. 13. № 10. С. 35-44.

13. Буслаев А.П., Кучелев Д.А., Яшина М.В. Динамические системы и математические модели трафика информации // Т-Comm: Телекоммуникации и транспорт. 2018. Т. 12. № 3. С. 22-38.

14. Miller A.J. Subset Selection in Regression. 2nd ed. Boca Raton: Chapman & Hall/CRC, 2002. 256 p.

15. Huber P.J., Ronchetti E.M. Robust Statistics. 2nd ed. Hoboken: Wiley, 2011. 384 p.

16. Rousseeuw P.J., Leroy A.M. Robust Regression and Outlier Detection. New York: Wiley, 1987. 352 p.

17. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning. 2nd ed. New York: Springer, 2021. 607 p.

18. Efron B., Hastie T., Johnstone I., Tibshirani R. Least Angle Regression // The Annals of Statistics. 2004. Vol. 32. No. 2, pp. 407-499.

19. Tibshirani R. Regression Shrinkage and Selection via the Lasso // Journal of the Royal Statistical Society: Series B. 1996. Vol. 58. No. 1, pp. 267-288.

20. Akaike H. A new look at the statistical model identification // IEEE Transactions on Automatic Control. 1974. Vol. 19. No. 6, pp. 716-723.

21. Kuhn M., Johnson K. Applied Predictive Modeling. New York: Springer, 2013. 600 p.

SELECTION OF INFORMATIVE REGRESSORS FOR MODELS CONSTRUCTED USING THE MAXIMUM CONSISTENCY CRITERION

Sergey I. Noskov, Irkutsk state transport university, Irkutsk, Russia, sergey.noskov.57@mail.ru

Abstract

This paper addresses the timely problem of selecting the most significant independent variables (regressors) in a linear regression model, whose parameters are estimated using the maximum consistency method (MCM). Unlike classical regression analysis, where model quality is traditionally assessed by minimizing the mean squared error, the sum of squared residuals, or other loss functions, MMC focuses on maximizing the number of observation pairs for which the signs of the increments of the observed and estimated values of the dependent variable coincide. This criterion, known as the consistency of behavior criterion (CBC), is highly robust to outliers and anomalous values, does not require assumptions about the normality of the error distribution, and allows for the reproduction of structural patterns in the data, including trend directions and inflection points. However, until now, there have been no regressor selection procedures adapted to the specifics of this discrete criterion, which has limited the application of MMS in problems with a large number of potential factors. This paper proposes, for the first time, an original algorithm for selecting regressors based on the sequential inclusion of variables that maximize the contribution to the increase in the value of the consistency criterion. This criterion, known as the consistency of behavior criterion (CBC), is highly robust to outliers and anomalous values, does not require assumptions about the normality of the error distribution, and allows for the reproduction of structural patterns in the data, including the direction of trends and points of trend reversal. However, until now, there have been no regressor selection procedures adapted to the specifics of this discrete criterion, which has limited the application of MMS in problems with a large number of potential factors. This paper proposes, for the first time, an original algorithm for selecting regressors based on the sequential inclusion of variables that maximize the contribution to the increase in the value of the consistency criterion. A step-by-step procedure for direct selection is described; for each candidate, a semi-integer linear programming problem is solved, allowing the maximum value of the consistency criterion to be precisely computed for a given set of regressors. Formal formulations of the problems are provided, and the solvability conditions (existence of a solution) and the monotonicity property of the criterion with respect to the inclusion of regressors are proven. An illustrative example using synthetic data, comprising 10 observations and 10 predictors, demonstrates the effectiveness of the proposed approach: the algorithm correctly identifies all significant variables and is robust to the addition of non-informative factors. The results can be used in building regression models for complex technical and transportation systems, in tasks involving the forecasting of telecommunications network load, the analysis of passenger flows, and other applied areas characterized by noisy data and the importance of correctly reproducing the direction of changes in the dependent variable.

Keywords: linear regression, feature selection, the method of maximum consistency, model specification, linear-Boolean programming

References

- [1] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Hoboken: Wiley, 2012.
- [2] Schoon E. W., Melamed D., Breiger R. L. *Regression Inside Out*. Cambridge: Cambridge University Press, 2024. 282 p.
- [3] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton: CRC Press LLC, 2015.
- [4] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1-22, 2010.
- [5] Breheny P., Huang J. *High-Dimensional Regression Modeling: Methodology, Applications, and Software*. Boca Raton, FL: CRC Press, 2024. 450 p.
- [6] Tibshirani R. J., Taylor J. "The solution path of the generalized lasso," *The Annals of Statistics*. Vol. 39, No. 3, pp. 1335-1371, 2011
- [7] S. I. Noskov, "Metod maksimal'noi soglasovannosti v regressionnom analize," *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki*, no. 10, pp. 380-385, 2021.
- [8] S. I. Noskov and Yu. A. Bychkov, "Vychislitel'nye eksperimenty s nepreryvnoi formoi metoda maksimal'noi soglasovannosti v regressionnom analize," *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta*, vol. 18, no. 2, pp. 7-12, 2022.
- [9] S. I. Noskov, "Kriterii 'soglasovannost povedeniya' v regressionnom analize," *Sovremennye tekhnologii. Sistemnyi analiz. Modelirovanie*, no. 1 (37), pp. 107-110, 2013.
- [10] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. New York: Springer, 2002.
- [11] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: Background, derivation, and applications," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, no. 2, pp. 199-203, 2012.
- [12] A.S. Bugaev, A.G. Tatashev, M.V. Yashina, O.S., Lavrov, E.A. Nosov, "Vosstanovlenie dinamiki transportnogo potoka na osnove determinirovanno-stokasticheskoy modeli i dannyy'x s intellektual'no transportny'x system," *T-Comm*, vol. 13., no. 10, pp. 35-44, 2019.
- [13] A.S. Bugaev, D.A. Kuchelelev, M.V. Yashina, "Dinamicheskie sistemy` i matematicheskie modeli trafika informacii," *T-Comm*, vol. 12, no. 3, pp. 22-38, 2018.
- [14] Miller A.J. *Subset Selection in Regression*. 2nd ed. Boca Raton: Chapman & Hall/CRC, 2002.
- [15] Huber P.J., Ronchetti E.M. *Robust Statistics*. 2nd ed. Hoboken: Wiley, 2009.
- [16] Rousseeuw P.J., Leroy A.M. *Robust Regression and Outlier Detection*. New York: Wiley, 1987.
- [17] James G., Witten D., Hastie T., Tibshirani R. *An Introduction to Statistical Learning*. 2nd ed. New York: Springer, 2021.
- [18] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, "Least Angle Regression," *The Annals of Statistics*, Vol. 32, No. 2. pp. 407-499, 2004.
- [19] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B*. Vol. 58, No. 1, pp. 267-288, 1996.
- [20] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 716-723, 1974.
- [21] M. Kuhn, K. Johnson, "Applied Predictive Modeling," New York: Springer, 2013.

Information about author:

Sergey I. Noskov, PhD, Professor, Irkutsk state transport university, Irkutsk, Russia. <https://orcid.org/0000-0003-4097-2720>