Математические методы исследования

УДК 519.2

ВЫБОР ОПТИМАЛЬНОГО НАБОРА ПРИЗНАКОВ ИЗ МУЛЬТИКОРРЕЛИРУЮЩЕГО МНОЖЕСТВА В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ[1]

© Р. Г. Нейчев[2], А. М. Катруца[2], В. В. Стрижов[3]

Статья поступила 9 июля 2015 г.

Рассмотрена проблема прогнозирования временных рядов. Для получения устойчивого прогноза предложено рассматривать входные временные ряды как матрицу объект-признак и использовать отбор признаков. В условиях мультиколлинеарности признаков необходим критерий для ее обнаружения. Для этого применяли подход, основанный на методе Белсли. Исключение коррелирующих признаков при отборе позволяет сократить размерность задачи и получить устойчивые оценки параметров модели. Для отбора признаков предложен метод добавления и удаления признаков. В качестве практической проверки данного метода в ходе вычислительного эксперимента решена задача прогнозирования почасовых значений цен на электроэнергию. Эксперименты проведены на реальных данных о ценах на электроэнергию в Германии.

Ключевые слова: устойчивость модели; выбор признаков; метод Белсли; почасовое прогнозирование цен; прогнозирование временных рядов; метод добавления-удаления признаков; линейная регрессия.

В современном мире большинство наблюдаемых величин, особенно связанных с деятельностью человека, обладают периодичностью: цены на авиабилеты значительно зависят от месяца; загруженность дорог — от дня недели и времени суток; спрос на сезонные товары — от времени года. В данной работе рассмотрена задача прогнозирования временных рядов, обладающих периодичностью. Для этого предлагается использовать метод авторегрессии [1], который сводит задачу прогнозирования к задаче линейной регрессии [2].

Кроме метода авторегрессии, для построения прогнозов используют другие подходы. Например, алгоритм *ARIMA*, являющийся обобщением алгоритма авторегрессионного скользящего среднего (*ARMA*), или алгоритм *Гусеницы* [3], который заключается в преобразовании одномерного временного ряда в многомерный и применении к нему метода главных компонент [4]. Сложность данного алгоритма квадратична по отношению к длине ряда, поэтому для длинных временных рядов он менее удобен, чем алгоритмы с линейной зависимостью. Под сложностью алгоритма понимается зависимость необходимых вычислительных ресурсов от длины подаваемого на вход временного ряда.

В задаче регрессии требуется восстановить значение целевого вектора на основе заданных признаков, наличие мультиколлинеарности между которыми приводит к получению в качестве решения неустойчивой модели. Модель будем называть *устойчивой*, если любые малые изменения вектора параметров приводят к слабым изменениям целевого вектора. *Мультиколлинеарность* — это наличие сильной зависимости между признаками, которая значительно снижает устойчивость модели [5]. Для решения проблемы мультиколлинеарности предлагается применить отбор признаков. Рассмотрим подходы, которые используют для отбора признаков. В методе *Lasso* [6] к среднеквадратичной ошибке добавляется регуляризационный член, равный l_1-норме вектора параметров. Это приводит к занулению некоторых элементов вектора параметров и, как следствие, к отбору признаков. Метод *LAD-Lasso* [6, 7] имеет преимущество перед *Lasso*, которое заключается в большей устойчивости к ошибкам и более точном отборе признаков благодаря

[1] Работа выполнена при финансовой поддержке РФФИ, проекты 14–07-31264, 16–07-01155.

[2] Московский физико-технический институт, Москва, Россия; e-mail: neychev@phystech.edu, amkatrutsa@yandex.ru

[3] Вычислительный центр РАН им. Дородницына, Москва, Россия; e-mail: strijov@gmail.com

тому, что каждый признак имеет собственную величину штрафа. Метод наименьших углов *LARS* [8] является обобщенной версией *Lasso*. Он оценивает веса свободных переменных, изменяя их таким образом, чтобы обеспечить наибольшую корреляцию с вектором регрессионных остатков. Основное преимущество этого метода заключается в том, что он выполняется за число шагов, не превышающее число свободных переменных. *Метод группового учета аргументов* [3] основан на рекурсивном отборе моделей, на основе которых строятся более сложные модели, за счет чего точность моделирования увеличивается на каждом шаге. *Гребневая регрессия* [10] применяется в случае мультиколлинеарности признаков, в ней в качестве регуляризатора к величине среднеквадратичной ошибки добавляется l_2-норма вектора параметров.

В данной работе в качестве способа отбора признаков предложена модифицированная версия шаговой регрессии *Add-Del* [1]. Добавление признаков проводится с помощью метода *FOS* [11]. Он последовательно добавляет признаки, максимально коррелирующие с вектором регрессионных остатков. Удаление признаков осуществляется с помощью *метода Белсли* [12], который позволяет обнаружить мультикоррелирующие признаки, исключив их из рассмотрения, и получить более устойчивую модель. Предполагается, что использование данных методов при отборе признаков позволит получить устойчивую модель, для работы с которой не требуется больших вычислительных ресурсов. Сложность предлагаемого метода линейно зависит от числа признаков.

Для практической проверки предлагаемого метода в рамках вычислительного эксперимента решается задача построения почасового прогноза цены на электроэнергию [13, 14]. Эксперимент поставлен на реальных данных и состоит из двух частей. В первой части на основе отобранных предлагаемым методом признаков строится прогноз методом авторегресси. Результаты сравниваются с реальным поведением цен после рассматриваемого периода и с данными работы других алгоритмов. Во второй части рассматривается непосредственно отбор признаков и сравниваются результаты работы предлагаемого и других методов.

Постановка задачи прогнозирования

Рассмотрим набор временных рядов $s_1, s_2, ..., s_p$ вида $s_j = \{x_{ij}\}$, $i = 1, ..., T-1$, $j = 1, ..., p$. Ряд s_1 будем называть *целевым*. Необходимо спрогнозировать следующие τ значений целевого ряда s_1, опираясь на ряды $s_1, ..., s_p$. Набор рядов обладает следующими свойствами: отсчеты времени i сделаны через равные промежутки, ряд s_j имеет периодическую составляющую τ и не имеет пропущенных значений, длина ряда $T-1$ кратна периоду τ.

Построим авторегрессионную матрицу для целевого ряда s_1:

$$\mathbf{X}_1 = \begin{pmatrix} x_1 & x_2 & \cdots & x_{\tau-1} & x_\tau \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{j\tau+1} & x_{j\tau+2} & \cdots & x_{(j+1)\tau-1} & x_{(j+1)\tau} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{(m-2)\tau+1} & x_{(m-2)\tau+2} & \cdots & x_{(m-1)\tau-1} & x_{(m-1)\tau} \\ x_{T-\tau+1} & x_{T-\tau+2} & \cdots & x_{T-1} & x_T \end{pmatrix}.$$

Представим матрицу $\mathbf{X}_1$ в виде

$$\mathbf{X}_1 = (\mathbf{X}|\mathbf{y}).$$

Для временных рядов s_j, $j = 2, ..., p$, построим авторегрессионные матрицы $\mathbf{X}_j$. Полученные матрицы припишем справа к матрице $\mathbf{X}$ для целевого ряда. В результате получим матрицу

$$\mathbf{X}^* = (\mathbf{X}|\mathbf{X}_2, ..., \mathbf{X}_p|\mathbf{y}).$$

Разобьем полученную авторегрессионную матрицу $\mathbf{X}^*$ следующим образом:

$$\mathbf{X}^* = \left(\mathbf{X}\mathbf{y}\mathbf{x}_m x_T\right),$$

выделив последний столбец и строку. Заметим, что вектор $\mathbf{y}$ и матрица $\mathbf{x}$ были переобозначены. Для нахождения значения x_T поставим задачу линейной регрессии:

$$||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 \to \min_w, \quad (1)$$

где $\mathbf{w}$ — вектор параметров. Тогда

$$x_T = \mathbf{x}_m\mathbf{w}. \quad (2)$$

Чтобы не перестраивать авторегрессионную матрицу для каждого следующего значения временного ряда, в качестве вектора x_m выбираются $\tau - 1$ значений ряда предыдущего периода. Необходимо найти такой вектор параметров $\mathbf{w}^*$, на котором достигается минимум функции ошибки S. Оптимальный вектор параметров для задачи (1) имеет вид

$$\mathbf{w}^* = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}(\mathbf{X}^\mathrm{T}\mathbf{y}).$$

При работе с временными рядами большой длины авторегрессионная матрица является недоопределенной. Например, рассмотрим временной ряд с почасовыми показаниями за два месяца. Матрица $\mathbf{X}$ будет иметь размеры $(24 - 1)(31 \cdot 2 - 1) = 23 \cdot 61$. Чтобы понизить ее размерность, поставим задачу отбора признаков. Признаками будем считать столбцы матрицы $\mathbf{X}$.

Постановка задачи отбора признаков

Представим целевой вектор в виде

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}(\mathbf{X}),$$

где $\boldsymbol{\varepsilon}$ — вектор регрессионных остатков. Пару $(\mathbf{X}, \mathbf{y})$ назовем выборкой и обозначим D. Признаки, которым

соответствуют ненулевые члены вектора **w**, назовем *активными*, а остальные признаки будем считать *исключенными*. Множество индексов элементов выборки обозначим I и разобьем на непересекающиеся подмножества: $J = L \cup C$.

Функцию ошибки $S(f(\mathbf{w}, \mathbf{X}), \mathbf{y})$ зададим как квадрат нормы вектора регрессионных остатков:

$$S = \|\boldsymbol{\varepsilon}(\mathbf{X})\|^2. \quad (3)$$

Подмножество индексов активных признаков обозначим $A \subset J$, где J — множество индексов всех признаков. Назовем моделью пару $(\mathbf{f}, A)$ и обозначим ее как $\mathbf{f}_A$. При постановке задачи линейной регрессии функция **f** фиксирована, поэтому для выбора модели необходимо найти множество индексов A, минимизирующее функцию ошибки S на элементах выборки D_C, состоящей из элементов выборки D с индексами из множества C:

$$A^* = \arg\min_{A \in J} S(A \,|\, \mathbf{w}^*, D_C). \quad (4)$$

Запись вида $S(A\,|D)$ означает, что выборка D фиксирована, а выборка A меняется. Для решения задачи (4) требуется найти вектор параметров $\mathbf{w}^*$, доставляющий минимум функции ошибки S на элементах выборки D_L, определяемой аналогично D:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in W} S(\mathbf{w}|D_L, A), \quad (5)$$

где W — множество возможных значений вектора **w**.

Метод Белсли

Между признаками возможно существование мультиколлинеарной зависимости. Напимер, если временной ряд обладает посуточной периодичностью, его значения в один и тот же час и в соседние часы сильно коррелируют. В случае мультиколлинеарности признаков оценка вектора параметров (5) является неустойчивой. Для устранения данной проблемы необходимо найти мультиколлинеарные признаки. Применим для этого метод Белсли. Рассмотрим сингулярное разложение матрицы **X**:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}, \quad (6)$$

где **U** и **V** — ортогональные матрицы; **D** — диагональная матрица, состоящая из сингулярных чисел λ_j таких, что $\lambda_1 < \lambda_2 < \ldots < \lambda_r$ (r — ранг матрицы **D**). Предполагаем, что матрица не вырождена, поэтому $r = n$. Ковариационная матрица вектора параметров **w** оценивается как $\hat{\mathbf{A}}^{-1} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$. Сингулярные числа λ_i являются собственными значениями, а столбцы матрицы **V** — собственными векторами ковариационной матрицы $\hat{\mathbf{A}}^{-1}$. Используя сингулярное разложение (6), запишем:

$$\hat{\mathbf{A}}^{-1} = \mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{V}\mathbf{D}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} = \mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{T}}.$$

Определим как i-й индекс обусловленности отношение $\eta_i = \lambda_{\max}/\lambda_i$, где $\lambda_{\max}$ — максимальное сингулярное число. Большое значение η_i указывает на близкую к линейной зависимость между признаками. Максимальный индекс обусловленности матрицы **X** показывает, насколько велико будет изменение компонент вектора параметров **w** при изменении матрицы признаков **X**. Назовем его числом обусловленности θ. Заметим, что число обусловленности матрицы $\hat{\mathbf{A}}$ есть квадрат числа обусловленности матрицы **X**.

В рамках нашей задачи для определения максимально коррелирующих признаков необходимо найти индекс

$$i^* = \arg\min_{i \in A} \eta_i, \quad (7)$$

где A — текущее множество активных признаков.

Оценками дисперсии параметров будут диагональные элементы матрицы $\hat{\mathbf{A}}^{-1}$. Дисперсионные доли q_{ij} определим как вклад j-го признака в дисперсию i-го элемента вектора параметров **w**:

$$q_{ij} = \frac{v_{ij}^2 \big/ \lambda_j^2}{\sum_{j=1}^{n} v_{ij}^2 \big/ \lambda_j^2}. \quad (8)$$

Из характеристики дисперсионных долей следует, что их большие значения указывают на наличие зависимости между признаками. Определим индекс j^*, вносящий максимальный вклад в дисперсию i-го элемента вектора **w**:

$$j^* = \arg\min_{j \in A} q_{i^* j}, \quad (9)$$

где максимальный индекс обусловленности i^* находится согласно (7).

Будем называть модель $\mathbf{f}_A$ неустойчивой, если число обусловленности матрицы признаков **X** велико: $\theta \gg 1$.

Исключение из модели $\mathbf{f}_A$ параметров, максимально влияющих на минимальное сингулярное значение матрицы **X**, обеспечивает максимальную устойчивость модели.

Действительно, пусть N объектов описываются моделью $\mathbf{f}_A$, состоящей из n признаков χ_i. Некоторые признаки мультиколлинеарны. Обозначим максимальный индекс обусловленности $\eta_{\max}$. Из ковариационной матрицы $\hat{\mathbf{A}}$ найдем максимальную соответствующую ему дисперсионную долю. Пусть она соответствует признаку χ_{j^*}. Это указывает на принадлежность признака χ_{j^*} к множеству мультикоррелирующих признаков. Мультиколлинеарность влечет близость к нулю одного или нескольких сингулярных значений λ_k.

Следовательно, исключение j^*-го признака из модели $\mathbf{f}_A$ приведет к увеличению минимального сингулярного числа и уменьшению числа обусловленности θ' новой матрицы $\mathbf{X}'$. Меньшее число обусловленно-

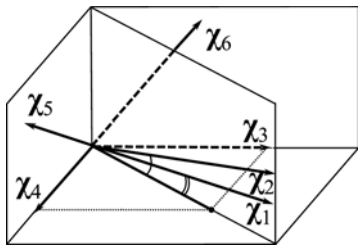

Рис. 1. Мультикоррелирующие ортогональные векторы значений свободных переменных

сти характеризует лучшую устойчивость модели, а значит, исключение j*-го признака ее повышает.

Для лучшего понимания принципа работы метода Белсли рассмотрим его на конкретном примере. На рис. 1 приведены три объекта, каждый из которых описывается шестью признаками χ_1, ..., χ_6. Расстоянием от объекта до признака будем считать величину угла между ними. Таблица 1 содержит разложение вектора дисперсий $\mathrm{var}(\chi_j)$ по индексам обусловленности η_i. На рис. 2 оценка ковариационной матрицы параметров $\hat{\mathbf{A}}$ визуализирована.

Согласно табл. 1 максимальный индекс обусловленности $\chi_6 = 4{,}87$. В соответствующей ему шестой строке дисперсионные доли в первом и втором столбцах (затемнены) максимальны. Они соответствуют признакам χ_1 и χ_2, между которыми имеется ярко выраженная зависимость (см. рис. 1).

Метод отбора признаков *Add-Del*

Применим *метод Белсли* для отбора признаков. Предлагаемый метод состоит из двух этапов — *Add* и *Del*. Метод предполагает, что регрессионные остатки имеют нормальное распределение с нулевым матожиданием и произвольной неотрицательно определенной ковариационной матрицей. В начальный момент множество активных признаков $A_0 = \varnothing$. Рассмотрим работу алгоритма.

Эman Add. Выбираем признак j* из множества исключенных признаков согласно

$$j^* = {}_{j \in J \setminus A} S(\mathbf{w}|A \cup \{j\}, D) \qquad (10)$$

и добавляем его к модели:

$$A = A \cup \{j^*\}. \qquad (11)$$

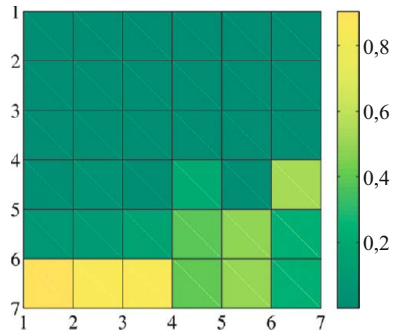

Рис. 2. Визуализация оценки ковариационной матрицы параметров $\hat{\mathbf{A}}$

Признаки добавляются к активному набору A до тех пор, пока функция ошибки не удовлетворит критерию останова

$$S(\mathbf{w}_A|D) \geq S_{\min} + \delta S_1. \qquad (12)$$

Эman Del. В соответствии с критерием (9) находим признак j*, максимально коррелирующий с другими, и удаляем его из множества:

$$A = A \setminus \{j^*\}. \qquad (13)$$

Повторяем данную операцию до тех пор, пока полученная ошибка $S(f_{A_k} | \mathbf{w}^*, D)$ превосходит свое минимальное значение на данном этапе не более чем на заданную величину δS_2. Критерий останова этапа *Del* на данном шаге:

$$S(\mathbf{w}_A|D_L) \geq S_{\min} + \delta S_2. \qquad (14)$$

Этапы *Add* и *Del* повторяются до тех пор, пока ошибка не стабилизируется, т.е. ее значение будет слабо изменяться от итерации к итерации. Колебание от итерации к итерации должно составлять величину порядка $1/100\,S$, более конкретное ее значение можно подобрать опытным путем. Остановимся подробней на критериях останова (12), (14). Для определения величин $\delta S_{1,2}$ рассмотрим свойства суммы квадратов регрессионных остатков

$$\mathbf{S} = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon}. \qquad (15)$$

Таблица 1. Разложение вектора дисперсий $\mathbf{var}(\chi_j)$ по индексам обусловленности η_i

η_i	Индекс обусловленности	Дисперсия параметров					
		$\mathbf{var}(\chi_1)$	$\mathbf{var}(\chi_2)$	$\mathbf{var}(\chi_3)$	$\mathbf{var}(\chi_4)$	$\mathbf{var}(\chi_5)$	$\mathbf{var}(\chi_6)$
η_1	0	0	0	0	0	0	0
η_2	0	0	0	0	0	0	0
η_3	0	0	0	0	0	0	0
η_4	1	0,02	0,05	0	0,2	0,01	0,53
η_5	2,29	0,08	0,1	0,15	0,39	0,48	0,23
η_6	4,87	0,9	0,85	0,84	0,4	0,5	0,24

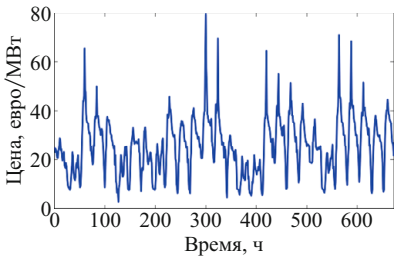

Рис. 3. Изменение цены в течение четырех недель

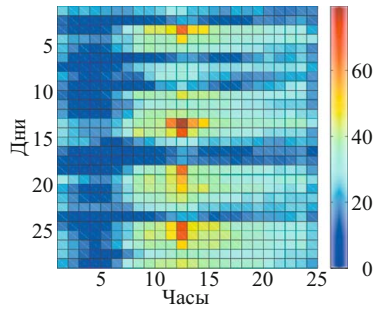

Рис. 4. Авторегрессионная матрица за четыре недели

Вектор регрессионных остатков

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\mathbf{w} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{y}, \qquad (16)$$

где матрица $(\mathbf{I} - \mathbf{X}(\mathbf{X}\mathbf{X})^{-1}\mathbf{X}) = \mathbf{P}$ симметрична и идемпотентна: $\mathbf{P} = \mathbf{P}$ и $\mathbf{P}^2 = \mathbf{P}$.

Подставляя выражение (16) в (15) и учитывая, что

$$\mathbf{w}^{\mathrm{T}} = \mathbf{y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1},$$

получаем

$$\mathbf{S} = \mathbf{y}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})\mathbf{X}^{\mathrm{T}})\mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \qquad (17)$$

Здесь $\mathbf{S}$ записана как квадратичная форма вектора $\mathbf{y}$. В предположении, что

$$\mathbf{y} \sim N(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}), \qquad (18)$$

математическое ожидание $\mathbf{S}$ будет иметь вид, как в [15].

Поскольку наиболее правдоподобная оценка параметров $\mathbf{w}$ при предположении (18) имеет вид $\mathbf{w} = (\mathbf{X}\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$ и функция ошибки в этом случае задана как $S(\mathbf{w}) = \mathbf{S}$, то математическое ожидание суммы квадратов регрессионных остатков составит

$$E(S(\mathbf{w})) = \mathrm{tr}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{I}\sigma^2 +$$

$$+ (\mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{X}\mathbf{w}. \qquad (19)$$

Так как след идемпотентной матрицы (в данном случае это матрица Мура – Пенроуза) $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}$, то

$$E(S(\mathbf{w})) = \mathrm{rank}((\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X})\mathbf{I}\sigma^2 =$$

$$= (\mathrm{m} - \mathrm{rank}(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}))\sigma^2 =$$

$$= (\mathrm{m} - \mathrm{rank}(\mathbf{X}))\sigma^2 = (\mathrm{m} - \mathrm{n})\sigma^2,$$

где m — число элементов выборки и строк матрицы $\mathbf{X}$. Если матрица плана $\mathbf{X}$ не содержит коллинеарных столбцов и ее ранг $\mathrm{rank}(\mathbf{X}) = n$, то несмещенной оценкой σ^2 является оценка

$$\sigma^2 = \frac{S(\mathbf{w})}{m-n}.$$

Поэтому величины $\delta S_{1,2}$ можно определить как дисперсию вектора регрессионных остатков $\boldsymbol{\varepsilon}$.

Вычислительный эксперимент

В рамках вычислительного эксперимента строили прогноз цен на электроэнергию. В ходе эксперимента сравнивали результаты прогнозирования с помощью предлагаемого метода (*Add-Del* и авторегрессия) и алгоритма *Гусеницы*. В качестве входных данных использовали временные ряды с реальными ценами на электроэнергию в Германии [16].

Для иллюстрации недельной и суточной периодичности на рис. 3 приведен график зависимости цены на электроэнергию от времени за четыре недели. На рис. 4 показана авторегрессионная матрица для того же периода. Видно, что цена имеет недельную и суточную периодичности, а значит, может иметь место мультиколлинеарность признаков, в роли которых в данной задаче выступают ежедневные значения цен.

На рис. 5 представлены прогнозы цены на пять суток с помощью авторегрессии с отбором признаков *Add-Del* (*а*) и *Гусеницы* (*б*). Прогноз, полученный с помощью предложенного метода, более точен при сопоставимом времени работы. Обучающая выборка содержит информацию за предшествующие 90 дней. Как видно, алгоритмы получают сходные по качеству результаты. Время работы алгоритмов практически одинаково.

Для оценки качества модели, полученной с помощью *Add-Del*, проводили ее сравнение с результатами методов *LARS* и *Lasso*, в которых также происходит отбор признаков путем обнуления коэффициентов вектора признаков $\mathbf{w}$. Для сравнения вычисляли информационные критерии Акаике *(AIC)* и Байеса *(BIC)*:

$$AIC = \tau\left(\ln\frac{RSS}{\tau}\right) + 2|A|,$$

$$BIC = \tau\left(\ln\frac{RSS}{\tau}\right) + |A|\ln\tau,$$

где RSS — величина среднеквадратичной ошибки соответствующего алгоритма, вычисленная по набору активных признаков A. При этом также учитывали число вошедших в модель признаков. Результаты сравнения приведены в табл. 2. На рис. 6 представле-

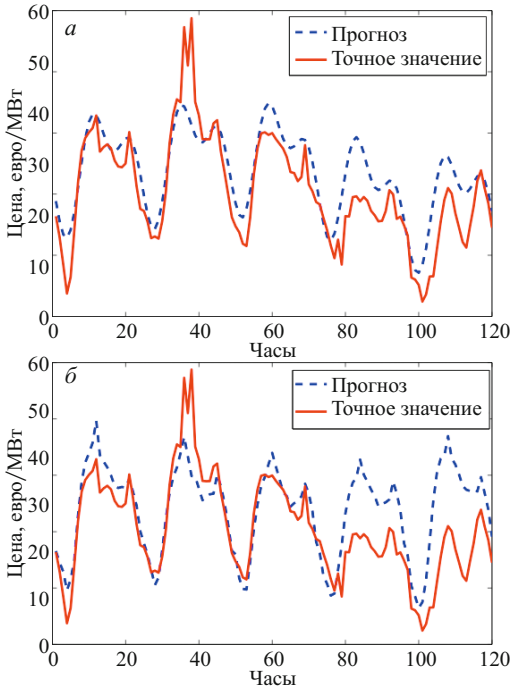

Рис. 5. Прогнозируемое и реальное поведение цены, полученные с использованием алгоритма *Add-Del* (*а*) и алгоритма *Гусеницы* (*б*)

ны результаты прогнозирования методом авторегрессии на пять недель с использованием полученных моделей.

Таким образом, рассмотрена проблема построения прогноза временных рядов. Подход, предлагаемый для ее решения, подразумевает исключение из рассмотрения некоторых входных данных для улучшения качества прогноза, т.е. отбор признаков. На устойчивость прогноза сильно влияет наличие мультиколлинеарных признаков, для обнаружения которых применяется метод Белсли. В ходе эксперимента оценены результаты работы предлагаемого метода, а также проведено сравнение качества моделей, полученных с помощью предлагаемого метода *Add-Del*, и алгоритмов *LARS* и *Lasso*. Предлагаемым методом получают модель, включающую меньшее количество признаков, но предоставляющую при этом сравнимые по качеству результаты. Область применения метода не ограничивается задачами прогнозирования, он может быть использован в любых задачах, требующих отбора признаков при построении модели по признаковому описанию.

ЛИТЕРАТУРА

1. **Леонтьева Л. Н.** Выбор моделей прогнозирования цен на электроэнергию / JMLDA. 2011. Т. 1. № 2. С. 127 – 137.
2. **Стрижов В. В., Крымова Е. А.** Алгоритм выбора признаков линейных регрессионных моделей из конечного и счетного множеств / Заводская лаборатория. Диагностика материалов. 2011. Т. 77. № 5. С. 63 – 68.
3. **Tsonis A. A., Elsner J. B.** Singular Spectrum Analysis. A New Tool in Time Series Analysis. — Springer US. 1996. — 164 p.

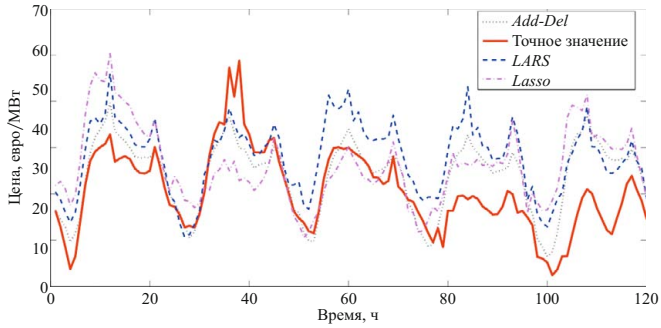

Рис. 6. Результаты работы алгоритмов, включающих отбор признаков

Таблица 2. Результаты построения модели

Метод	*RSS*	*AIC*	*BIC*	Число признаков *n*
Add-Del	57,97	237,97	343,99	11
LARS	82,19	178,82	283,67	23
Lasso	104,01	282,01	386,86	18

REFERENCES

1. **Leontyeva L. N.** Vybor modelei prognozirovaniya tsen na élektroénergiyu / JMLDA. 2011. Vol. 1. N 2. P. 127 – 137 [in Russian].
2. **Krymova E. A., Strizhov V. V.** Algoritm vybora priznakov lineinykh regressionnykh modelei iz konechnogo i schetnogo mnozhestv [Selection algorithm for linear regression models of finite and countable sets] / Zavod. Lab. Diagn. Mater. 2011. Vol. 77. N 5. P. 63 – 68.
3. **Tsonis A. A., Elsner J. B.** Singular Spectrum Analysis. A New Tool in Time Series Analysis. — Springer US. 1996. — 164 p.
4. **Zinovyev A. Y., Gorban A. N., Sumner N. R.** Topological grammars for data approximation / Appl. Math. Lett. 2007. Vol. 20. N 4. P. 382 – 386.
5. **Chi-Hyuck Jun, Il-Gyo Chong**. Performance of some variable selection methods when multicollinearity is present / Chemometrics and Intelligent Laboratory Systems. 2005. Vol. 78. N 1, 2. P. 103 – 112.
6. **Jiang Guohua, Wang Hansheng, Li Guodong**. Robust regression shrinkage and consistent variable selection through the LAD-lasso / J. Business Econ. Stat. 2008. Vol. 25. P. 347 – 355.
7. **Herzog F., Hildmann M.** Robust calculation and parameter estimation of the hourly price forward curve / 17th Power Systems Computation Conference. Stockholm. 2011. P. 1 – 7.
8. **Efron B., Hastie T., Johnstone I., Tibshirani R.** Least angle regression / The Annals of Statistics. 2004. Vol. 32. N 3. P. 407 – 499.
9. **Степашко В. С., Ивахненко А. Г.** Помехоустойчивость моделирования. — Киев: Наукова думка, 1985. — 216 с.
10. **Smith H., Draper N. R.** Appied regression analysis. — New York: John Wihley and Sons, 1998. — 736 p.
11. **Grant P. M., Chen S., Cowan S. F. N.** Orthogonal least squares learning algorithm for radial basis function network / Neural Networks. 1991. Vol. 2. N 2. P. 302 – 309.
12. **Belsley A. D.** Conditioning Diagnostics: Collinearity and Weak Data in Regression. — New York: John Wiley and Sons, 1991. — 396 p.
13. **Abdolkhalig A.** Optimized calculation of hourly price forward curve (HPFC) / Int. J. Electr. Comp. Electronics Comm. Eng. 2008. Vol. 2. N 9. P. 840 – 850.
14. **Caro G., Hildmann M.** What makes a good hourly price forward curve? / European Energy Market, IEEE 10th International Conference, 2013. Stockholm. P. 1 – 7.
15. **Kachapova F., Kachapov I.** Orthogonal projection in teaching regression and financial mathematics / J. Stat. Education. 2010. Vol. 18. N 1. P. 1 – 18.
16. Временной ряд цен на электроэнергию: https://svn.code.sf.net/p/dmba/code/data/germanspotprice.csv

4. **Zinovyev A. Y., Gorban A. N., Sumner N. R.** Topological grammars for data approximation / Appl. Math. Lett. 2007. Vol. 20. N 4. P. 382 – 386.
5. **Chi-Hyuck Jun, Il-Gyo Chong**. Performance of some variable selection methods when multicollinearity is present / Chemometrics and Intelligent Laboratory Systems. 2005. Vol. 78. N 1, 2. P. 103 – 112.
6. **Jiang Guohua, Wang Hansheng, Li Guodong**. Robust regression shrinkage and consistent variable selection through the LAD-lasso / J. Business Econ. Stat. 2008. Vol. 25. P. 347 – 355.
7. **Herzog F., Hildmann M.** Robust calculation and parameter estimation of the hourly price forward curve / 17th Power Systems Computation Conference. Stockholm. 2011. P. 1 – 7.
8. **Efron B., Hastie T., Johnstone I., Tibshirani R.** Least angle regression / The Annals of Statistics. 2004. Vol. 32. N 3. P. 407 – 499.
9. **Stepashko V. S., Ivakhnenko A. G.** Pomekhoustoichivost' modelirovaniya [Noise Immunity of modeling]. — Kiev: Naukova dumka, 1985. — 216 p. [in Russian].
10. **Smith H., Draper N. R.** Appied regression analysis. — New York: John Wihley and Sons, 1998. — 736 p.
11. **Grant P. M., Chen S., Cowan S. F. N.** Orthogonal least squares learning algorithm for radial basis function network / Neural Networks. 1991. Vol. 2. N 2. P. 302 – 309.
12. **Belsley A. D.** Conditioning Diagnostics: Collinearity and Weak Data in Regression. — New York: John Wiley and Sons, 1991. — 396 p.
13. **Abdolkhalig A.** Optimized calculation of hourly price forward curve (HPFC) / Int. J. Electr. Comp. Electronics Comm. Eng. 2008. Vol. 2. N 9. P. 840 – 850.
14. **Caro G., Hildmann M.** What makes a good hourly price forward curve? / European Energy Market, IEEE 10th International Conference, 2013. Stockholm. P. 1 – 7.
15. **Kachapova F., Kachapov I.** Orthogonal projection in teaching regression and financial mathematics / J. Stat. Education. 2010. Vol. 18. N 1. P. 1 – 18.
16. Time series with electricity prices: https://svn.code.sf.net/p/dmba/code/data/germanspotprice.csv

УДК 519.24

ЛОКАЛЬНАЯ АСИМПТОТИЧЕСКАЯ НОРМАЛЬНОСТЬ СТАТИСТИЧЕСКИХ ЭКСПЕРИМЕНТОВ И ЕЕ РОЛЬ В ТЕОРИИ ОЦЕНИВАНИЯ И ПРОВЕРКИ ГИПОТЕЗ

© **А. А. Абдушукуров, Н. С. Нурмухамедова**[1]

Статья поступила 18 июня 2015 г.

Основной задачей теории оценивания является нахождение оптимальных оценок для неизвестных параметров. Существуют два подхода к решению этих задач. Первый основан на выборке конечного объема, второй — асимптотический — на выборке с растущим объемом. Асимптотический подход может обладать свойствами оптимальности при $n \to \infty$. Он базируется на понятии асимптотической минимаксности оценок. Локальная асимптотическая минимаксность оценок опирается на асимптотическое поведение последовательности статистических экспериментов при сближающихся последовательностях альтернативных гипотез. В данной работе рассмотрена асимптотическая нормальность оценок байесовского типа и асимптотически минимаксная эффективность оценок максимального правдоподобия с использованием свойства локальной асимптотической нормальности статистики отношения правдоподобия в модели случайного цензурирования с двух сторон.

Ключевые слова: локальная асимптотическая нормальность; статистика отношения правдоподобия; асимптотическая минимаксная эффективность; случайное цензурирование.

Статистика отношения правдоподобия (СОП) играет фундаментальную роль в теории принятия решений, особенно в теории проверки статистических гипотез. Пусть $(\mathbf{X}^{(n)}, \mathbf{B}^{(n)}, \mathbf{P}^{(n)})$ — статистическая модель, соответствующая повторной независимой выборке наблюдений $X^{(n)} = (X_1, ..., X_n)$ случайной величины (с.в.) X с распределением $Pl_\theta \in \mathbf{P} = \{P_\theta, \theta \in \Theta\}$, где Θ — открытое множество в R^1, θ — неизвестный параметр. При общей постановке задач теории проверки гипотез предполагается, что неизвестный параметр исходного распределения P_θ принадлежит заданному подмножеству $\Theta_0 \in \Theta$ (гипотеза H_0) множества возможных значений параметра θ. Дополнительное предположение (альтернативная гипотеза к основной гипотезе H_0) подразумевает, что $\theta \in \Theta_1 = \Theta \backslash \Theta_0$ (гипотеза H_1). Основная задача теории состоит в проверке соответствия реальных экспериментальных данных предполагаемой гипотезе на основе статистического критерия, т.е. процедуры, позволяющей принимать или отвергать данную гипотезу. Критерии дают возможность утверждать, что результаты наблюдений не противоречат принятой гипотезе, т.е. статистические выводы формулируются в следующем виде: экспериментальные данные согласуются с данной гипотезой (или

[1] Национальный университет Узбекистана им. М. Улугбека, г. Ташкент, Узбекистан; e-mail: a_abdushukurov@rambler.ru