УДК 519.584

А. М. Катруца, студент, Московский физико-технический институт,
В. В. Стрижов, доктор физ.-мат. наук, науч. сотр.,
Вычислительный центр РАН имени А. А. Дородницина, Москва, e-mail: strijov@gmail.com

Проблема мультиколлинеарности при выборе признаков в регрессионных задачах

Исследуется проблема мультиколлинеарности и ее влияние на эффективность методов выбора признаков. Предлагается процедура тестирования методов выбора признаков и методика порождения тестовых выборок с различными типами мультиколлинеарности между признаками. Рассматриваемые методы выбора признаков тестируются на порожденных выборках. Процедура тестирования заключается в применении методов выбора признаков к выборкам с различным типом мультиколлинеарности и оценивании количества мультиколлинеарных признаков в множестве отобранных признаков. В работе приводится критерий сравнения методов выбора признаков. Методы выбора признаков сравниваются согласно различным функционалам качества. Проведено сравнение методов выбора признаков для случая наличия в данных определенного типа мультиколлинеарности. Сделан вывод о качестве работы рассматриваемых методов на определенных типах данных.

Ключевые слова: *регрессионный анализ, выбор признаков, мультиколлинеарность, тестовые выборки, критерий качества*

Введение

Работа посвящена тестированию методов выбора признаков. Предполагается, что исследуемая выборка содержит значительное число мультиколлинеарных признаков. *Мультиколлинеарность* — это сильная корреляционная связь между отбираемыми для анализа признаками, совместно воздействующими на целевой вектор, которая затрудняет оценивание регрессионных параметров и выявление зависимости между признаками и целевым вектором. Проблема мультиколлинеарности, возможные способы ее обнаружения и устранения описаны в работах [1—3]. Также мультиколлинеарность приводит к уменьшению устойчивости оценок вектора параметров. Оценка вектора параметров называется устойчивой, если малое изменение некоторой компоненты этого вектора приводит к малому изменению соответствующей компоненты оценки целевого вектора.

В задачах анализа данных для уменьшения размерности [4, 5], упрощения использования стандартных алгоритмов машинного обучения [6], удаления нерелевантных признаков [7] и повышения обобщающей способности применяемого алгоритма [8] применяют методы выбора признаков. Также методы выбора признаков используют для решения проблемы мультиколлинеарности в задачах регрессии [9].

Задача выбора оптимального подмножества признаков является одной из основных задач предварительной обработки данных. Методы выбора признаков основаны на минимизации некоторого функционала, который отражает качество рассматриваемого подмножества признаков. В работах [10—12] сделан обзор существующих методов выбора признаков, проведена классификация методов выбора признаков по используемым функционалам качества и стратегии поиска оптимального подмножества признаков.

При наличии мультиколлинеарности в регрессионных задачах применение методов выбора признаков приводит к повышению устойчивости оценок параметров и уменьшению их дисперсии. Для этого применяют методы отбора признаков с различными регуляризаторами или стратегиями добавления и удаления признаков с использованием статистических тестов для проверки значимости добавляемого признака. Примерами методов, использующих регуляризаторы, являются гребневая регрессия [13], где регуляризатор — взвешенная евклидова норма вектора параметров; Lasso [14] и LARS [15], где регуляризатор — взвешенная сумма модулей параметров; Elastic net [16], где регуляризатор — линейная комбинация предыдущих двух регуляризаторов. Методом, использующим проверку значимости добавляемого или удаляемого признака, является шаговая регрессия [17] с различными ком-

бинациями процедур добавления или удаления признаков.

Для тестирования методов выбора признаков в работе [9] предложен метод генерации выборок и функционал, позволяющий оценить качество процедуры выбора признаков. Однако предложенный способ не позволяет оценить изменение критерия качества при непрерывном изменении параметров выборок и структурного параметра мультиколлинеарности.

В нашей работе предложена другая процедура генерации тестовых выборок, основанная на задании свойств признаков. Рассматриваются следующие свойства признаков: мультиколлинеарность между признаками; коррелированность целевому вектору; ортогональность между признаками; ортогональность признаков целевому вектору. Задание количества признаков, обладающих каждым из этих свойств, позволяет генерировать выборки с различным взаимным расположением признаков и целевого вектора. Такой метод генерации тестовых выборок дает возможность исследовать зависимость эффективности методов выбора признаков при непрерывном изменении параметра мультиколлинеарности.

В работе предложен критерий ранжирования методов выбора признаков и методика их тестирования. Критерием ранжирования является число мультиколлинеарных признаков в множестве отобранных признаков, удаление которых приводит к росту ошибки не больше некоторого заданного значения. Методика тестирования заключается в последовательном применении различных методов выбора признаков к тестовым выборкам, каждая из которых отражает некоторый тип мультиколлинеарности, и оценке качества полученного подмножества признаков для каждой пары, включающей метод выбора признаков и тестовую выборку.

1. Постановка задачи выбора признаков

Задана выборка $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in I = \{1, ..., m\}$, множество свободных переменных — вектор $\mathbf{x} = [x_1, ..., x_j, ..., x_n]$, где $j \in J = \{1, ..., n\}$. Предполагается, что эти переменные принадлежат множеству действительных чисел, либо его подмножеству: $x_i \in \mathbb{X} \subseteq \mathbb{R}^n$ и $y_i \in \mathbb{Y} \subseteq \mathbb{R}^1$. Введем обозначения: $\mathbf{y} = [y_1, ..., y_m]^\mathrm{T}$ — вектор значений зависимой переменной, целевой вектор; $\boldsymbol{\chi}_j = [x_{1j}, ..., x_{mj}]^\mathrm{T}$ — реализация j-й свободной переменной; j-й признак и $\mathbf{X} = [\mathbf{x}_1^\mathrm{T}, ..., \mathbf{x}_n^\mathrm{T}]^\mathrm{T} = [\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n]$ — матрица плана. Предполагается, что вектор $\mathbf{x}_i$ и число y_i связаны соотношением

$$y_i = f(\mathbf{w}, \mathbf{x}_i) + \varepsilon(\mathbf{x}_i), \qquad (1)$$

где f:$\mathbb{W} \times \mathbb{X} \to \mathbb{Y}$ отображение декартова произведения пространства допустимых параметров $\mathbb{W}$ и пространства значений $\mathbb{X}$ свободной переменной в область значений $\mathbb{Y}$ зависимой переменной, а $\varepsilon(\mathbf{x}_i)$ — i-й компонент вектора регрессионных остатков $\boldsymbol{\varepsilon} = \mathbf{f} - \mathbf{y}$. Обозначим вектор-функцию

$$\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{X}) = [f(\mathbf{w}, \mathbf{x}_1), ..., f(\mathbf{w}, \mathbf{x}_m)]^\mathrm{T} \in \mathbb{Y}^m.$$

Определим функцию ошибки

$$S\colon \mathbb{X} \times \mathbb{W} \times \mathbb{Y} \to \mathbb{R}_+$$

и представление множества индексов элементов выборки в виде

$$I = L \cup C.$$

Далее в качестве функции ошибки S зададим квадрат нормы вектора регрессионных остатков $\boldsymbol{\varepsilon}$:

$$S = \sum_{i=1}^{m} \varepsilon^2(\mathbf{x}_i) = \mathrm{RSS};\ \mathrm{TSS} = \sum_{i=1}^{m} (y_i - \overline{y})^2, \qquad (2)$$

где $\overline{y} = \frac{1}{m} \sum_{i=1}^{m} y_i$; RSS (residual sum of squares) — сумма квадратов регрессионных остатков; TSS (total sum of squares) — полная сумма квадратов.

Требуется найти такой оптимальный вектор параметров $\mathbf{w}^* \in \mathbb{W}$, при котором функция приближает целевой вектор $\mathbf{y}$ наилучшим образом в смысле функции ошибки S.

Назовем моделью пару $(\mathbf{f}, A)$, где $A \subset J$ — подмножество индексов признаков, используемое для вычисления вектор-функции $\mathbf{f}$. Ниже фиксирована функция $\mathbf{f} = \mathbf{X}\mathbf{w}$, после этого модель зависит только от множества A, поэтому вместо $(\mathbf{f}, A)$ для обозначения применяемой модели будем использовать A. Таким образом, выбор модели сводится к нахождению оптимального множества индексов A^* в смысле функции ошибки S, вычисляемой на элементах выборки D_C:

$$A^* = \arg\min_{A \subset J} S(A|\mathbf{w}^*, D_C). \qquad (3)$$

Для решения задачи (3) необходимо найти вектор параметров $\mathbf{w}^*$ как решение задачи минимизации функции ошибки на элементах выборки D_L с индексами из множества L:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \subset \mathbb{W}} S(\mathbf{w}|A, D_L). \qquad (4)$$

Задача (3) является задачей выбора признаков и заключается в нахождении подмножества индексов признаков $A^* \subset J$, минимизирующего функцию ошибки S.

2. Анализ мультиколлинеарности при выборе признаков

В дальнейшем будем считать, что векторы признаков $\boldsymbol{\chi}_j$ и целевой вектор $\mathbf{y}$ нормированы. Рассмотрим некоторое подмножество $B \subset J$ индексов признаков. Назовем признаки *мультиколлинеарными*,

если найдутся такие коэффициенты a_k, $k \in B$ и достаточно малое $\delta > 0$, что

$$\|\boldsymbol{\chi}_j - \sum_{k \in B} a_k \boldsymbol{\chi}_k\| < \delta, \quad (5)$$

где j — индекс признака и $j \notin B$. Чем меньше δ, тем выше степень мультиколлинеарности.

Назовем признаки с индексами i, j *коррелирующими*, если найдется достаточно малое δ_{ij} такое, что

$$\|\boldsymbol{\chi}_j - \boldsymbol{\chi}_i\| < \delta_{ij}. \quad (6)$$

Из определения следует, что и формула (6) есть частный случай формулы (5) при $\delta_{ji} = \delta_{ij}$ и $a_k = 1$, $k = j$.

Назовем признак $\boldsymbol{\chi}_j$ *коррелированным с целевым вектором*, если найдется достаточно малое δ_{yj}, такое что

$$\|\mathbf{y} - \boldsymbol{\chi}_j\| < \delta_{yj}.$$

2.1. Фактор инфляции дисперсии

Широкоизвестным критерием анализа мультиколлинеарности авторы считают фактор инфляции дисперсии [18]. Фактор инфляции дисперсии VIF_j определяется для j-го признака и является показателем наличия линейной зависимости между j-м и остальными признаками. Для нахождения VIF_j необходимо определить оценку $\hat{\mathbf{w}}$ для вектора коэффициентов $\mathbf{w}$ в задаче (1) при $y_i = x_{ij}$, $i \in I$ и $J = J \setminus j$. Аналогично (2) определяются RSS и TSS. Величина VIF_j определяется следующим выражением:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}, \quad (7)$$

где $R_j^2 = 1 - \dfrac{\mathrm{RSS}}{\mathrm{TSS}}$ — коэффициент детерминации. Согласно [18] значение $\mathrm{VIF}_j \gtrsim 5$ означает наличие зависимости между j-м и всеми остальные признаками. Недостатками этого критерия мультиколлинеарности является то, что он может принимать большие значения сразу для нескольких признаков, что мешает определить какой из признаков необходимо удалить.

Другим критерием наличия мультиколлинеарности между признаками является число обусловленности κ матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, которое равно отношению максимального и минимального по модулю собственных чисел $\lambda_{\max}$ и $\lambda_{\min}$:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}. \quad (8)$$

Оно показывает насколько матрица близка к вырожденной. Чем больше κ, тем ближе матрица к вырожденной.

2.2. Метод Белсли

Для обнаружения и исключения мультиколлинеарных признаков в наборе отобранных признаков предлагается явно поставить оптимизационную задачу, используя метод Белсли. Критерием сравнения методов выбора признаков в данной работе является критерий, основанный на исключении признака, мультиколлинеарного некоторым другим признакам из набора выбранных признаков. Исключение проводится методом Белсли. Предлагаемый критерий сравнения методов выбора признаков в дальнейшем называется критерием наличия мультиколлинеарных признаков среди отобранных признаков. Будем считать, что на множестве параметров $\mathbb{W}$ задано нормальное распределение

$$\mathbf{w} \sim N(\mathbf{w}_{ML}, \mathbf{A}^{-1})$$

с матожиданием $\mathbf{w}_{ML}$ и ковариационной матрицей $\mathbf{A}^{-1}$. Оценка $\hat{\mathbf{A}}^{-1}$ ковариационной матрицы в случае линейной модели будет

$$\hat{\mathbf{A}}^{-1} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}.$$

Используя сингулярное разложение матрицы $\mathbf{X}$

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}},$$

где $\mathbf{U}$ и $\mathbf{V}$ — ортогональные матрицы, а $\boldsymbol{\Lambda}$ — диагональная с собственными значениями λ_i на диагонали, такими что

$$\lambda_1 > \lambda_2 > \ldots > \lambda_n,$$

получим выражение для $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-2}\mathbf{V}^{-1}.$$

Столбцы матрицы $\mathbf{V}$ — собственные векторы, а квадраты сингулярных чисел — собственные значения матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}} = \mathbf{V}\boldsymbol{\Lambda}^2\mathbf{V}^{\mathrm{T}},$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}^2.$$

Отношение максимального собственного значения $\lambda_{\max}$ к i-му собственному значению λ_i назовем индексом обусловленности η_i:

$$\eta_i = \frac{\lambda_{\max}}{\lambda_i}.$$

Большое значение η_i указывает на зависимость, близкую к линейной, между признаками и чем больше η_i, тем сильнее зависимость. Поэтому на этапе удаления нужно найти такой индекс i^*, что

$$i^* = \arg\max_{i \in F_{k-1}} \eta_i,$$

где F_{k-1} текущее подмножество признаков. Оценками дисперсий параметров будут диагональные элементы матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$:

$$\mathrm{Var}(w_i) = \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_j^2}.$$

Далее определим дисперсионную долю q_{ij} как вклад j-го признака в дисперсию i-го элемента вектора параметров $\mathbf{w}$:

$$q_{ij} = \frac{v_{ij}^2/\lambda_j^2}{\sum_{j=1}^{n} v_{ij}^2/\lambda_j^2},$$

где $[v_{ij}] = \mathbf{V}$, а λ_j — собственное значение. Большие значения дисперсионных долей означают наличие зависимостей между признаками, это следует из способа их получения.

Следовательно, по найденному максимальному индексу обусловленности i^* находим признак j^*:

$$j^* = \arg\min_{i \in F_{k-1}} q_{i^*j}, \qquad (9)$$

который вносит наибольший вклад в дисперсию i-го элемента вектора $\mathbf{w}$, т. е. является коллинеарным некоторому другому признаку.

3. Методы построения тестовых выборок

Для тестирования методов выбора признаков предлагается использовать синтетические выборки, которые определяются с помощью множеств P_f, P_y, C_f, C_y и R, индексирующих соответственно ортогональные признаки, признаки ортогональные целевому вектору, мультиколлинеарные признаки, признаки, коррелирующие с целевым вектором и случайно расположенные признаки. Определим следующие множества, задающие структуру выборки:

1) множество ортогональных признаков $\boldsymbol{\chi}_j$ с индексами j из множества P_f;

2) множество признаков $\boldsymbol{\chi}_j$, ортогональных целевому вектору $\mathbf{y}$, с индексами j из множества P_y;

3) множество мультиколлинеарных признаков $\boldsymbol{\chi}_j$ с индексами j из множества C_f;

4) множество признаков $\boldsymbol{\chi}_j$, коррелирующих с целевым вектором, с индексами j из множества C_y;

5) множество случайных признаков $\boldsymbol{\chi}_j$ с индексами из множества R.

Для регулирования степени мультиколлинеарности используется параметр мультиколлинеарности k: при $k = 1$ признаки коллинеарны, при $k = 0$ — ортогональны.

При этом параметр k используется как для определения степени мультиколлинеарности признаков, так и для определения степени коррелированности признаков и целевого вектора.

Рассмотрим базовые варианты взаимного расположения мультиколлинеарных признаков и целевого вектора, из которых варьированием параметров можно генерировать различные выборки для тестирования методов выбора признаков.

1. Признаки $\boldsymbol{\chi}_j$ с индексами как из множества мультиколлинеарных между собой признаков $j \in C_f$, так и из множества ортогональных целевому вектору $\mathbf{y}$, $j \in P_y$:

$$\langle \mathbf{y}, \boldsymbol{\chi}_j \rangle = 0, j \in J; \|\boldsymbol{\chi}_i - \sum_{l \in B} a_l \boldsymbol{\chi}_l\| < \delta, i \in J, i \notin B \subset J; \qquad (10)$$

$$J = P_y \cap C_f.$$

Схематично взаимное расположение признаков и целевого вектора изображено на рис. 1. Выборки с такой структурой будем называть выборками первого типа.

2. Все признаки $\boldsymbol{\chi}_j$ порождены случайно из многомерной случайной величины, $j \in R$. Эта случайная величина взята из равномерного распределения на единичном кубе размерности r. При этом найдется некоторый признак $\boldsymbol{\chi}_i$, $i \in R$, приближающий целевой вектор $\mathbf{y}$:

$$J = R, |R| = r, \boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r \sim U[0, 1]^r, \|\mathbf{y} - \boldsymbol{\chi}_i\| < \delta. \qquad (11)$$

Схематично взаимное расположение признаков и целевого вектора изображено на рис. 2. Выборки с такой структурой будем называть выборками второго типа.

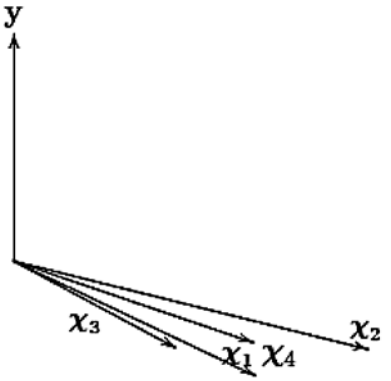


Рис. 1. Неадекватная коррелирующая выборка

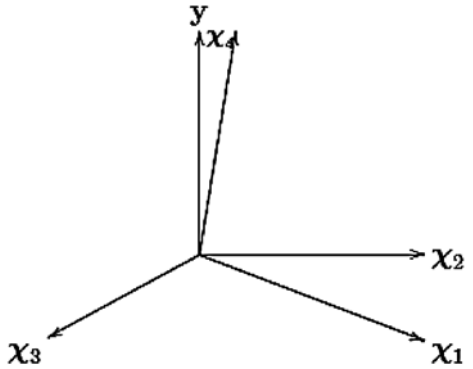


Рис. 2. Адекватная случайная выборка

3. Все признаки $\boldsymbol{\chi}_j$ коррелируют и хорошо приближают целевой вектор $\mathbf{y}$:

$$\|\boldsymbol{\chi}_j - \boldsymbol{\chi}_i\| < \delta_{ij},\ i, j \in J;\ \|\mathbf{y} - \boldsymbol{\chi}_j\| < \delta,\ j \in J,\ J = C_y. \quad (12)$$

Схематично расположение признаков и целевого вектора изображено на рис. 3. Выборки с такой структурой будем называть выборками третьего типа.

4. Множество признаков $\boldsymbol{\chi}_j$ с индексами из множества $j \in J$ состоит из объединения двух множеств: множества ортогональных признаков с индексами из множества P_f и множества признаков $\boldsymbol{\chi}_c$, коррелированных с некоторыми из них. Индексы c лежат в множестве C_f. При этом целевой вектор $\mathbf{y}$ хорошо приближается линейной комбинацией ортогональных признаков $\boldsymbol{\chi}_j$, $j \in J$:

$$\langle \boldsymbol{\chi}_i, \boldsymbol{\chi}_j \rangle = 0,\ i, j \in P_f;\ \mathbf{y} = \sum_{j \in P_f} a_j \boldsymbol{\chi}_j;$$
$$\|\boldsymbol{\chi}_j - \boldsymbol{\chi}_i\| < \delta_{ij},\ i \in P_f, j \in C_f;\ J = P_f \cup C_f. \quad (13)$$

Схематично взаимное расположение признаков и целевого вектора изображено на рис. 4. Выборки с такой структурой будем называть выборками четвертого типа.

Комбинируя описанные выше варианты взаимного расположения признаков и целевого вектора, варьируя параметр мультиколлинеарности, а также изменяя мощности p_f, p_y, c_f, c_y и r множеств P_f, P_y, C_f, C_y и R, индексирующих множества признаков со свойствами, определенными в выражениях (10)—(13), можно генерировать выборки для тестирования методов выбора признаков.

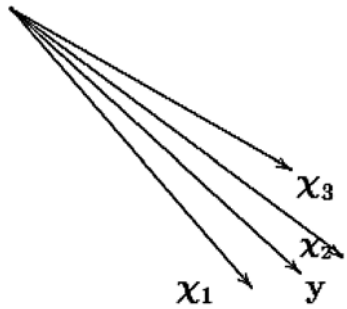


Рис. 3. Адекватная избыточная выборка

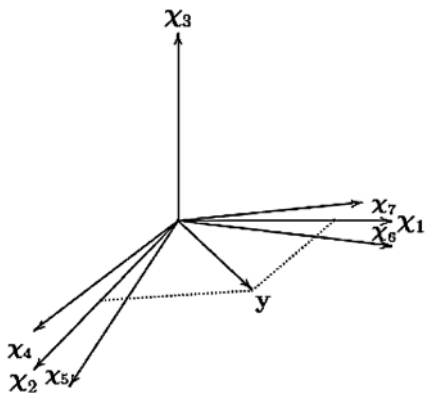


Рис. 4. Адекватная коррелирующая выборка

4. Критерии сравнения методов выбора признаков

Для анализа методов выбора признаков определим следующий критерий, позволяющий оценить сколько мультиколлинеарных признаков есть в множестве отобранных признаков. Зададим некоторое предельное значение s_0 функции ошибки S. Результатом работы метода выбора признаков является набор признаков с индексами из множества $A \subset J$, $p = |A|$. Для найденного множества признаков получен оптимальный вектор параметров $\mathbf{w}_A^*$. Назовем h максимальную мощность множества индексов признаков $J_h \subset A$, при удалении которого значение функции ошибки S не превосходит s_0:

$$h = \arg\max_{S(J_h, \mathbf{w}_h, D) < s_0} |J_h|, \quad (14)$$

где $S(J_h, \mathbf{w}_h, D)$ — функция ошибки, в которой первый аргумент — это матрица $\mathbf{X}$ со столбцами, индексы которых лежат в множестве J_h, второй аргумент — вектор параметров $\mathbf{w}_h$, составленный из элементов $\mathbf{w}_A^*$ с индексами из множества J_h, и третий аргумент — выборка. Ниже в разделе "Вычислительный эксперимент" определялась величина d, значение которой равно числу признаков, удаление их приводит к ошибке, не превышающей s_0:

$$d = |A| - h. \quad (15)$$

Определение индексов удаляемых признаков проводилось методом Белсли, задача (9). Методы выбора признаков ранжируются по возрастанию величины d: большие значения d показывают, что выбранное подмножество признаков (решение задачи (3)) содержит избыточные признаки, удаление которых приводит к росту функции ошибки вплоть до s_0.

Ранее авторами [18, 19] были предложены следующие критерии сравнения линейных регрессионных моделей.

1. Скорректированный (adjusted) коэффициент детерминации R^2_{adj} учитывает добавление избыточных признаков и выражается как

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(m-p)}{\text{TSS}/(m-1)}, \quad (16)$$

где $p = |A|$, m — число строк в матрице $\mathbf{X}$, а RSS и TSS определяются из (2). Чем ближе значение к единице, тем лучше модель описывает целевой вектор.

2. Критерий C_p позволяет достичь компромисса между значением RSS и числом используемых переменных $p = |A|$, а также ликвидировать возможную коллинеарность признаков. Величина C_p определяется следующим образом:

$$C_p = \frac{\text{RSS}_A}{\text{RSS}} - m + 2p, \quad (17)$$

где RSS_A — это величина, аналогичная RSS, но найденная при использовании признаков с индексами из множества A. Меньшие значения соответствуют лучшему набору признаков.

3. Информационный критерий BIC вычисляется по следующей формуле:

$$\mathrm{BIC} = \mathrm{RSS} + p\log m, \qquad (18)$$

где $p = |A|$ и m — это количество строк в матрице $\mathbf{X}$. Чем меньше величина BIC, тем лучше модель описывает целевой вектор.

4. F-тест используется в случае линейной модели для проверки отсутствия релевантных признаков. Если ни один из признаков не приближает целевой вектор, то величина

$$\frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1} \qquad (19)$$

имеет распределение Фишера с p, $n - p - 1$ степенями свободы.

5. Вычислительный эксперимент

В вычислительном эксперименте проведено сравнение методов выбора признаков по различным функционалам качества при фиксированном значении предельной функции ошибки $s_0 = 0{,}5$ и при двух значениях параметра мультиколлинеарности $k = 0{,}2$ и $k = 0{,}8$. Для каждой выборки и для каждого метода выбора признаков были получены зависимости между предельным значением функции ошибки s_0 и максимальным числом d, а также между фактором инфляции дисперсии VIF и параметром мультиколлинеарности k. При этом VIF определялся для признаков из множества A, что показывает наличие мультиколлинеарных признаков в множестве отобранных признаков A. Эксперименты проводили на выборках при $k = 0{,}2$ и $k = 0{,}8$. Для выборок второго типа график зависимости VIF от параметра мультиколлинеарности k и числа избыточных признаков d в множестве отобранных признаков от предельного значения функции ошибки s_0 не строили, так как в этом типе выборок нет мультиколлинеарных признаков.

Результаты экспериментов сведены в табл. 1—7.

В экспериментах генерировались выборки четырех типов, определяемых формулами (10)—(13) для двух значений параметра мультиколлинеарности $k = 0{,}2$ и $k = 0{,}8$. Перед проведением экспериментов векторы признаков и целевой вектор были отнормированы, так что евклидова норма векторов признаков и целевого вектора равна единице. Измеряемые значения критериев усреднены по пяти повторениям. Значения элементов вектора $\mathbf{w}$, меньшие 10^{-6}, считались незначительными и равными нулю. Значения p-value соответствуют проверке нулевой гипотезы о том, что вектор параметров $\mathbf{w}$ — нулевой, т. е. отсутствуют признаки, с помощью которых можно приблизить целевой вектор $\mathbf{y}$, против альтернативы, что среди столбцов матрицы $\mathbf{X}$ есть подходящие для описания целевого вектора $\mathbf{y}$, при уровне значимости 0,05. Если значение p-value меньше 0,05, то нулевая гипотеза отвергается. Это означает, что среди признаков есть такие, которые хорошо приближают целевой вектор $\mathbf{y}$. Проверка выполняется с помощью F-теста (19). В таблицах, где отсутствует столбец со значениями p-value, они пренебрежимо малы. Значение предельной функции ошибки $s_0 = 0{,}5$.

Сравнивали методы LARS, Lasso, ElasticNet, Ridge и Stepwise. Все, кроме последнего, являются методами, которые одновременно решают задачи (4) и (3). Отбор признаков проводится обнулением незначащих коэффициентов в оптимальном векторе пара-

Таблица 1

Значения функционалов качества для выборок первого типа при $k = 0{,}2$

Метод	d	C_p	RSS	к	VIF	R^2_{adj}	BIC	p-value
Lasso	0	−997	1	3,84	1,05	−3,32	314,62	0,11
Ridge	0	−997	1	4,13	1,05	−3,31	346,39	0,1
LARS	—	−997	—	—	—	—	—	—
Stepwise	0	−997	1	4,13	1,05	−3,41	346,41	$5{,}28 \cdot 10^{-4}$
Elastic Net	0	−997	1	3,84	1,05	−3,32	314,32	0,11

Таблица 2

Значения функционалов качества для выборок первого типа при $k = 0{,}8$

Метод	d	C_p	RSS	к	VIF	R^2_{adj}	BIC	p-value
Lasso	0	−997	1	717,8	16,6	−3,32	310,48	0,06
Ridge	0	−997	1	801	16,6	−3,31	346,39	0,05
LARS	—	−997	—	—	—	—	—	—
Stepwise	0	−997	1,68	801	16,6	−6,22	347,01	10^{-10}
Elastic Net	0	−997	1	717,8	16,6	−3,32	310,48	0,06

Таблица 3

Значения функционалов качества для выборок второго типа

Метод	d	C_p	RSS	к	VIF	R^2_{adj}	BIC
Lasso	0	$7 \cdot 10^6$	$8{,}50 \cdot 10^{-4}$	1	0,25	1	6,9
Elastic Net	0	$8{,}76 \cdot 10^{-4}$	$8{,}76 \cdot 10^{-4}$	1	0,25	1	6,9
Ridge	0	$7{,}97 \cdot 10^9$	0,97	1	0,25	−3	7,88
LARS	0,2	−997	$1{,}30 \cdot 10^{-10}$	2,19	0,32	1	8,29
Stepwise	4,6	−997	$1{,}33 \cdot 10^{-10}$	28,86	0,89	1	53,88

Таблица 4

Значения функционалов качества для выборок третьего типа при $k = 0{,}2$

Метод	d	C_p	RSS	$\text{к} \cdot 10^8$	$\text{VIF} \cdot 10^7$	R^2_{adj}	BIC
Ridge	0	$2{,}3 \cdot 10^9$	0,97	24	1,14	−3,17	346,36
Lasso	1	$2 \cdot 10^6$	$8{,}50 \cdot 10^{-4}$	0,95	0,58	1	13,82
Elastic Net	3,2	$2 \cdot 10^6$	$8{,}50 \cdot 10^{-4}$	2,8	0,97	1	41,45
LARS	36	−997	$4{,}22 \cdot 10^{-10}$	24	1,14	1	345,39
Stepwise	36	−997	$4{,}22 \cdot 10^{-10}$	24	1,14	1	345,39

Таблица 5

Значения функционалов качества для выборок третьего типа при $k = 0{,}8$

Метод	d	C_p	RSS	к	VIF	R^2_{adj}	BIC
Lasso	0	$5{,}16 \cdot 10^8$	$8{,}50 \cdot 10^{-4}$	1	0,24	1	6,9
Ridge	0	$5{,}9 \cdot 10^{11}$	0,97	$6{,}07 \cdot 10^{11}$	$2{,}9 \cdot 10^9$	−3,17	346,36
Elastic Net	3,2	$5{,}16 \cdot 10^8$	$8{,}50 \cdot 10^{-4}$	$7{,}3 \cdot 10^{10}$	$2{,}5 \cdot 10^9$	1	41,45
LARS	36	−997	$1{,}65 \cdot 10^{-12}$	$6{,}07 \cdot 10^{11}$	$2{,}9 \cdot 10^9$	1	345,39
Stepwise	36	−997	$1{,}73 \cdot 10^{-12}$	$6{,}07 \cdot 10^{11}$	$2{,}9 \cdot 10^9$	1	345,39

Таблица 6

Значения функционалов качества для выборок четвёртого типа при $k = 0{,}2$

Метод	d	C_p	RSS	к	VIF	R^2_{adj}	BIC
Ridge	0	$6 \cdot 10^{30}$	0,95	$8{,}42 \cdot 10^{15}$	$1{,}15 \cdot 10^{23}$	−3	210,95
Stepwise	1	−868,95	$5{,}45 \cdot 10^{-29}$	1	0,63	1	13,82
LARS	1,8	$5{,}38 \cdot 10^{29}$	0,38	$2{,}1 \cdot 10^{16}$	$3{,}3 \cdot 10^{30}$	−0,62	102,62
Elastic Net	17,6	$5{,}84 \cdot 10^{27}$	$9{,}18 \cdot 10^{-4}$	$1{,}4 \cdot 10^{16}$	$5{,}32 \cdot 10^{20}$	1	150,59
Lasso	18	$5{,}84 \cdot 10^{27}$	$9{,}18 \cdot 10^{-4}$	$1{,}4 \cdot 10^{16}$	$5{,}32 \cdot 10^{20}$	1	150,60

Таблица 7

Значения функционалов качества для выборок четвёртого типа при $k = 0{,}8$

Метод	d	C_p	RSS	к	VIF	R^2_{adj}	BIC
Ridge	0	$1{,}8 \cdot 10^{30}$	0,95	10^{16}	$8{,}65 \cdot 10^{16}$	−2,97	152,92
Stepwise	1	$9{,}4 \cdot 10^5$	$8{,}8 \cdot 10^{-25}$	1	0,63	1	13,82
LARS	1,2	10^{30}	0,38	$3 \cdot 10^{29}$	10^{20}	−0,57	108,15
Lasso	14,8	$1{,}73 \cdot 10^{27}$	$9{,}2 \cdot 10^{-4}$	$9{,}92 \cdot 10^{15}$	10^{17}	1	150,59
Elastic Net	15,2	$1{,}70 \cdot 10^{27}$	$9{,}2 \cdot 10^{-4}$	$9{,}92 \cdot 10^{15}$	10^{17}	1	150,59

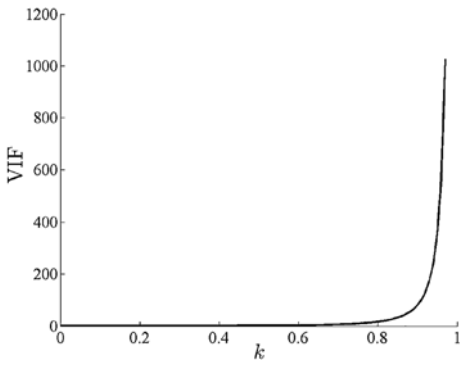
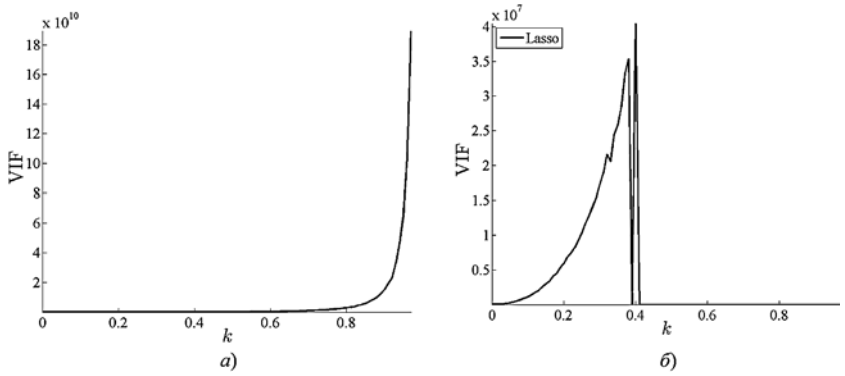


Рис. 5. Зависимость фактора инфляции дисперсии VIF от параметра мультиколлинеарности *k* для первого типа выборок



а) *б*)

Рис. 6. Зависимость фактора инфляции дисперсии VIF от параметра мультиколлинеарности *k* для третьего типа выборок:
а — при работе всех рассматриваемых методов отбора кроме Lasso; *б* — при работе метода Lasso

метров **w***. Метод Stepwise последовательно решает задачи (3) и (4), добавляя и удаляя признаки в соответствии с их значимостью, определяемой статистическим тестом. В алгоритме ElasticNet используется взвешенная сумма регуляризаторов из алгоритмов Lasso и Ridge, веса у обоих регуляризаторов равны 0,5. Прочерк в таблице означает, что метод выбора признаков не отбирает ни один признак и получаемый вектор **w*** нулевой.

В столбцах таблиц стоят ранее введенные критерии качества модели: число мультиколлинеарных признаков d в множестве отобранных признаков (15); критерий C_p (17); остаточная сумма квадратов RSS (2); число обусловленности κ матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ (8); значение VIF (7); скорректированный коэффициент детерминации R^2_{adj} (16); информационный критерий BIC (18) и p-value для F-теста (19).

Для выборок первого типа (10) $n = p_y = 50$, $m = 1000$ результаты приведены в табл. 1 и 2 при $k = 0{,}2$ и $k = 0{,}8$ соответственно.

Для выборок второго типа (11) $n = r = 50$, $m = 1000$ результаты приведены в табл. 3.

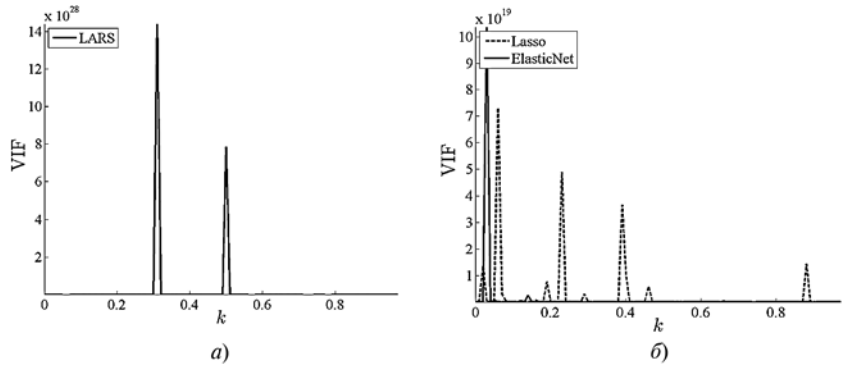
Для выборок третьего типа (12) $n = c_y = 50$, $m = 1000$ результаты приведены в табл. 4 и 5 при $k = 0{,}2$ и $k = 0{,}8$ соответственно.

Для выборок четвертого типа (13) $p_f = 10$, $c_f = 40$, $m = 1000$ результаты приведены в табл. 6 и 7 при $k = 0{,}2$ и $k = 0{,}8$ соответственно.

На рис. 5—8, представлена зависимость VIF от параметра мультиколлинеарности k для каждого типа выборок, где эта зависимость имеет место.

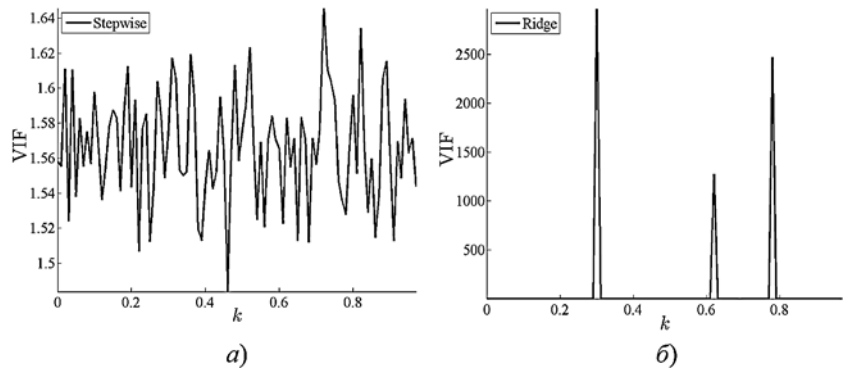
На рис. 5 показана зависимость VIF от параметра мультиколлинеарности k для первого типа выборок при работе различных алгоритмов. На этом рисунке видно, что все алгоритмы показывают одинаковые результаты, и ни один из рассматриваемых методов выбора признаков не решает проблему мультиколлинеарности в случае ортогональности всех признаков целевому вектору и взаимной коррелированности.

На рис. 6 изображена зависимость VIF от параметра мультиколлинеарности k для третьего типа выборок. Видно, что все методы показывают одинаковый вид зависимости, кроме метода Lasso. Для него при росте параметра мультиколлинеарности наблю-



а) *б*)

Рис. 7. Зависимость фактора инфляции дисперсии VIF от параметра мультиколлинеарности *k* для четвертого типа выборок:
а — при работе метода LARS; *б* — при работе методов Lasso и Elastic Net



а) *б*)

Рис. 8. Зависимость фактора инфляции дисперсии VIF от параметра мультиколлинеарности *k* для четвертого типа выборок:
а — при работе метода Stepwise; *б* — при работе метода Ridge

дается резкое уменьшение значения величины VIF. Это говорит об отсутствии линейной зависимости между выбранными признаками в выборках, сгенерированных при $k \gtrsim 0{,}4$.

На рис. 7, 8 показана зависимость VIF от параметра мультиколлинеарности k для четвертого типа выборок при работе различных методов. Метод LARS показывает резкие скачки значений VIF, как и метод Ridge, но у метода Ridge амплитуда скачков ниже. Методы Lasso и ElasticNet демонстрируют скачки, схожие со скачками у LARS, но меньшей амплитуды и более высокой частоты. Для выборок четвертого типа после применения метода Stepwise значения VIF не превышают двух при росте коэффициента k. Это означает, что метод Stepwise для выборок четвертого типа дает набор линейно независимых признаков.

Рассмотрим зависимость числа мультиколлинеарных признаков d в множестве отобранных признаков от значений предельной ошибки s_0 для ранее рассмотренных типов выборок на рис. 9—11.

На рис. 9 показана зависимость числа лишних признаков d в множестве отобранных признаков от предельного значения функции ошибки s_0 для первого типа выборок при значениях $k = 0{,}2$ и $k = 0{,}8$. Значение d стабильно равно нулю вследствие ортогональности целевого вектора и всех признаков вплоть до значений, близких к единице. Далее идет резкий скачок d, так как предельное значение функции ошибки выросло достаточно, чтобы удалить сразу почти все признаки.

На рис. 10 показана зависимость величины d от параметра s_0 для третьего типа выборок при значении $k = 0{,}2$ и $k = 0{,}8$. Метод Lasso отбирает один

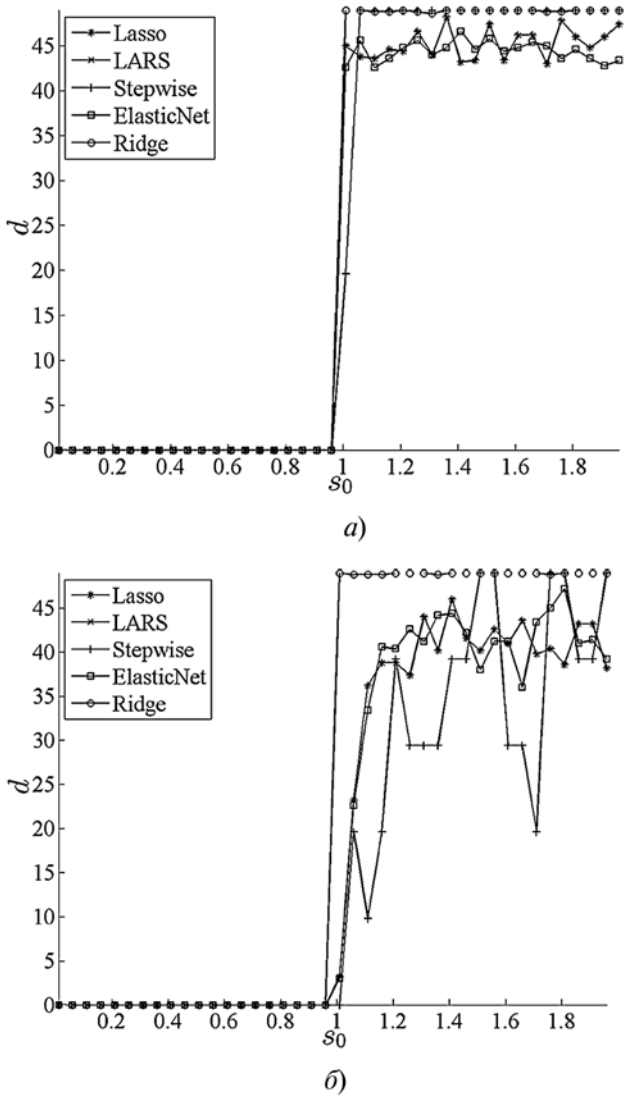


Рис. 9. Зависимость числа мультиколлинеарных признаков d в множестве отобранных признаков от предельного значения функции ошибки s_0 для первого типа выборок:
а — при $k = 0{,}2$; *б* — при $k = 0{,}8$

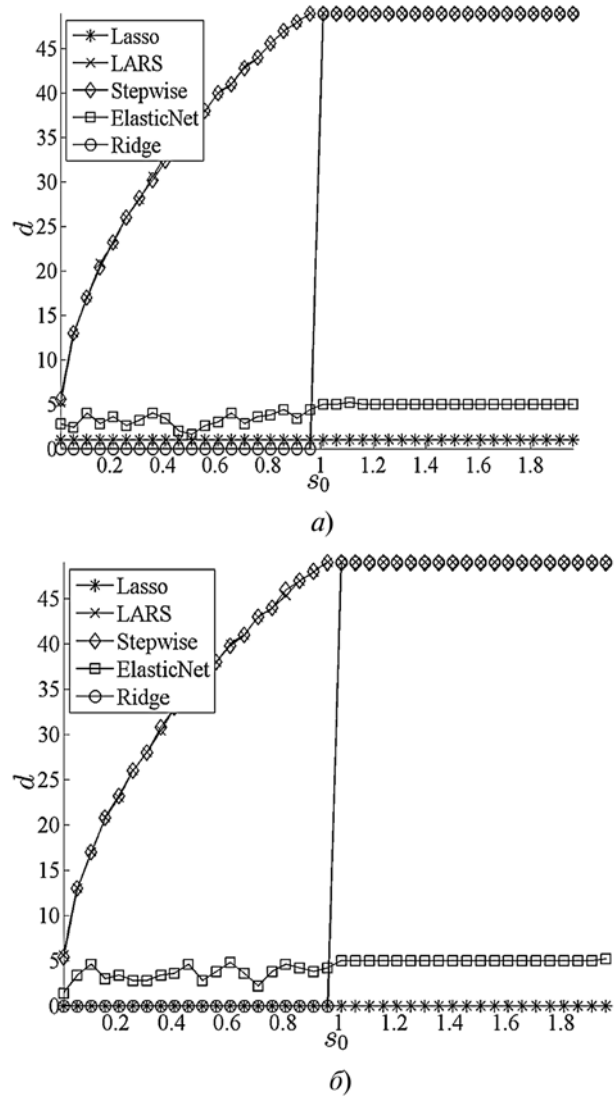


Рис. 10. Зависимость числа мультиколлинеарных признаков d в множестве отобранных признаков от предельного значения функции ошибки s_0 для третьего типа выборок:
а — при $k = 0{,}2$; *б* — при $k = 0{,}8$

или два признака, наилучшим образом приближающие целевой вектор, поэтому значение d для этого метода равно нулю или единице. Аналогично, но чуть хуже, работает метод Elastic Net, он отбирает чуть больше лишних признаков нежели метод Lasso. Зависимость d от s_0 для метода Ridge схожа с зависимостью для первого типа выборок по той же причине: сначала она достаточна велика, чтобы удалять хоть одни признак, но как только приближается к единице становится возможным удалить сразу почти все признаки. Для методов LARS и Stepwise наблюдается постепенный рост величины d с ростом предельного значения функции ошибки s_0 с выходом на константу при достижении s_0 значения, близкого к единице.

На рис. 11 показана зависимость d от параметра s_0 для четвертого типа выборок при $k = 0{,}2$ и $k = 0{,}8$.

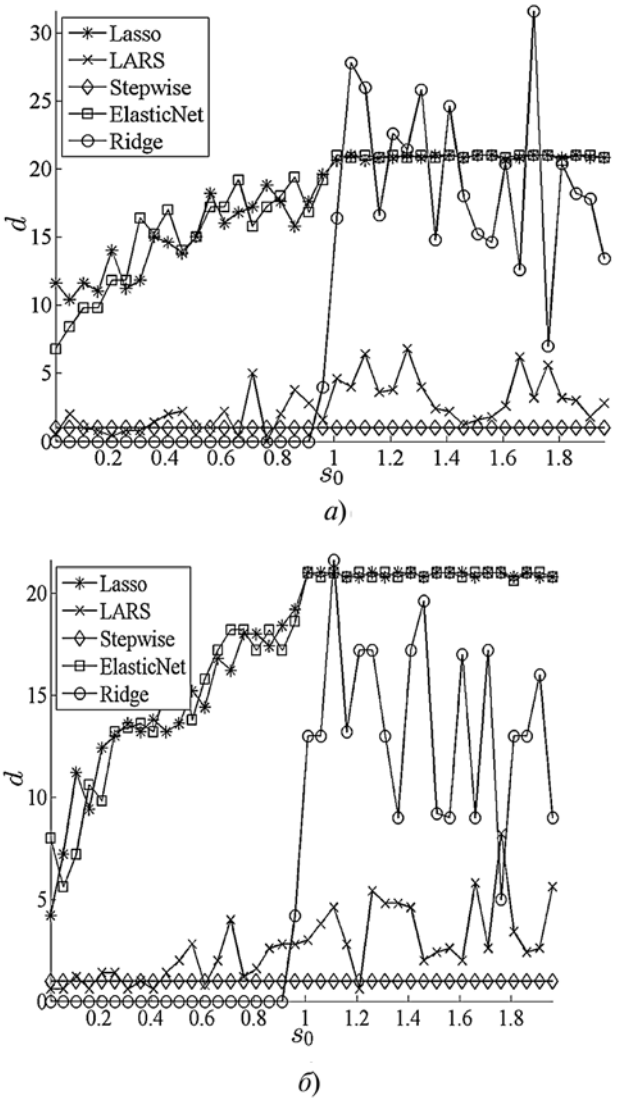


Рис. 11. Зависимость числа мультиколлинеарных признаков d в множестве отобранных признаков от предельного значения функции ошибки s_0 для четвертого типа выборок:
а — при $k = 0{,}2$; *б* — при $k = 0{,}8$

Наиболее стабильные решения дает метод Stepwise, у которого в среднем обнаруживается только один признак, удаление которого приводит к ошибке, не превышающей s_0. Чуть хуже работает метод LARS: число лишних признаков d среди отобранных им не превышает пяти при росте предельного значения функции ошибки s_0. Для методов Lasso и ElasticNet наблюдается рост d при росте s_0 до единицы, а затем стабилизация на уровне $d \simeq 20$. Для метода Ridge вид зависимости схож с предыдущими типами выборок, только для четвертого типа после преодоления s_0 значения, равного единице, величина d начала сильно колебаться. Это показывает неустойчивость набора признаков, получаемого методом Ridge для четвертого типа выборок.

Заключение

В работе проведено исследование эффективности методов выбора признаков в случае выборок с мультиколлинеарными признаками. Эксперименты показали, что из рассмотренных методов проблему мультиколлинеарности при отборе признаков решают методы Lasso (для выборок третьего типа) и Stepwise (для выборок четвертого типа). Для выборок первого типа все рассмотренные методы показывают одинаковые результаты: ни один из рассматриваемых методов выбора признаков не решает проблему мультиколлинеарности в случае ортогональности всех признаков целевому вектору. Предложенный критерий показывает, что как при малых, так и при больших значениях k устойчивые решения дают одинаковые методы. Также вид зависимости между величинами s_0 и d практически одинаков в рамках одной выборки для больших и маленьких значений k. Для выборок первого типа все рассматриваемые методы показывают одинаковый результат, для выборок третьего типа наиболее устойчивый результат дает метод Lasso, для выборок четвертого типа — методы LARS и Stepwise.

Работа выполнена при поддержке РФФИ, проект 14-07-31046.

Список литературы

1. **Askin R. G.** Multicollinearity in regression: Review and examples // Journal of Forecasting. 1982. V. 1, N. 3. P. 281—292.

2. **Learner E. E.** Multicollinearity: A Bayesian Interpretation // The Review of Economics and Statistics. 1973. V. 55, N. 3. P. 371—380.

3. **Belsley D. A., Kuh E., Welsch R. E.** Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons, 2005.

4. **Yu Lei, Liu Huan.** Feature selection for high-dimensional data: A fast correlation-based filter solution // ICML. Washington D. C. 2003. V. 3. P. 856—863.

5. **Стрижов В. В., Кузнецов М. П., Рудаков К. В.** Метрическая кластеризация последовательностей аминокислотных остатков в ранговых шкалах // Математическая биология и биоинформатика. 2012. Т. 7. № 1. С. 345—359.

6. **Chen Yi-Wei, Lin Chih-Jen.** Combining SVMs with various feature selection strategies // Feature Extraction. Foundations and Applications. Berlin: Springer, 2006. P. 315—324.

7. **George H. J., Kohavi R., Pfleger K.** et al. Irrelevant Features and the Subset Selection Problem // ICML. New Brunswick. 1994. V. 94. P. 121—129.

8. **Vorontsov K.** Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. 2008. V. 18, N. 2. P. 243—259.

9. **Chong Il-Gyo, Jun Chi-Hyuck.** Performance of some variable selection methods when multicollinearity is present // Chemometrics and Intelligent Laboratory Systems. 2005. V. 78, N. 1—2. P. 103—112.

10. **Guyon I., Elisseeff A.** An Introduction to Variable and Feature Selection // The Journal of Machine Learning Research. 2003. V. 3. P. 1157—1182.

11. **Bolón-Canedo V., Sánchez-Maroño N., Alonso-Betanzos A.** A review of feature selection methods on synthetic data // Knowledge and information systems. 2013. V. 34, N. 3. P. 483—519.

12. **Ladha L., Deepa T.** Feature Selection Methods and Algorithms // International Journal on Computer Science & Engineering. 2011. V. 3, N. 5.

13. **El-Dereny M., Rashwan N. I.** Solving multicollinearity problem using ridge regression models // International Journal of Contemporary Mathematical Sciences. 2011. V. 6. P. 585—600.

14. **Tibshirani R.** Regression Shrinkage and Selection Via the Lasso // Journal of the Royal Statistical Society, Series B. — 1994. — V. 58. — P. 267—288.

15. **Efron B., Hastie T., Tibshirani R.** Least angle regression // The Annals of tatistics. 2004. V. 32, N. 2. P. 407—499.

16. **Zou Hui, Hastie T.** Regularization and variable selection via the Elastic Net // Journal of the Royal Statistical Society, Series B. — 2005. — V. 67. — P. 301—320.

17. **Harrell F. E.** Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. — Berlin: Springer, 2001.

18. **Paul R. K.** Multicollinearity: causes, effects and remedies // Accessed. Apr. 23, 2013. URL: http://bit.ly/lqDHObV.

19. **Strijov V., Krymova E., Weber G.-W.** Evidence optimization for consequently generated models. // Mathematical and Computer Modeling. 2013. V. 57, N. 1—2. P. 50—56.

A. M. Katrutsa, Student,
Moscow Institute of Physrcs and Technology, Moscow, amkatrutsa@yandex.ru,
V. V. Strijov, Rechecher,
Dorodnicyn Computing Center of Russian Academy of Sciences, Moscow, strijov@ccas.com

The Multicollinearity Problem for Feature Selection Methods in Regression

The paper investigates the multicollinearity problem in regression analysis and its influence on the performance of feature selection methods. The authors propose a procedure to test feature selection methods. A criteria is proposed to compare the feature selection methods, according to their performance when the multicollinearity is present. The feature selection methods are compared according to the other well-known evaluation measures. Methods to generate data sets of different multicollinearity types were proposed. The authors investigate performance of feature selection methods. The feature selection methods were tested on the data sets of different multicollinearity types.

Keywords: *regression analysis, feature selection, multicollinearity, test data sets*

References

1. **Askin R. G.** Multicollinearity in regression: Review and examples. *Journal of Forecasting.* 1982. V. 1, N. 3. P. 281—292.

2. **Learner E. E.** Multicollinearity: A Bayesian Interpretation. *The Review of Economics and Statistics.* 1973. V. 55, N. 3. P. 371—380.

3. **Belsley D. A., Kuh E., Welsch R. E.** *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: John Wiley & Sons, 2005.

4. **Yu Lei, Liu Huan.** Feature selection for high-dimensional data: A fast correlation-based filter solution. *ICML.* Washington D. C., 2003. V. 3. P. 856—863.

5. **Strijov V. V., Kuznetsov M. P., Rudakov K. V.** Metricheskaya klasterizatsiya posledovatel'nostey aminokislotnykh ostatkov v rangovykh shkalakh. *Matematicheskaya biologiya i bioinformatika.* 2012. V. 7, N. 1. P. 345—359.

6. **Chen Yi-Wei, Lin Chih-Jen.** Combining SVMs with various feature selection strategies. *Feature Extraction. Foundations and Applications.* Berlin: Springer, 2006. P. 315—324.

7. **George H. J., Kohavi R., Pfleger K.** et al. Irrelevant Features and the Subset Selection Problem. *ICML.* New Brunswick, 1994. V. 94. P. 121—129.

8. **Vorontsov K.** Combinatorial probability and the tightness of generalization bounds. *Pattern Recognition and Image Analysis.* 2008. V. 18, N. 2. P. 243—259.

9. **Chong Il-Gyo, Jun Chi-Hyuck.** Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems.* 2005. V. 78, N. 1—2. P. 103—112.

10. **Guyon I., Elisseeff A.** An Introduction to Variable and Feature Selection. *The Journal of Machine Learning Research.* 2003. V. 3. P. 1157—1182.

11. **Bolón-Canedo V., Sánchez-Maroño N., Alonso-Betanzos A.** A review of feature selection methods on synthetic data. *Knowledge and information systems.* 2013. V. 34, N. 3. P. 483—519.

12. **Ladha L., Deepa T.** Feature Selection Methods and Algorithms. *International Journal on Computer Science & Engineering.* V. 3, N. 5. 2011.

13. **El-Dereny M., Rashwan N. I.** Solving multicollinearity problem using ridge regression models. *International Journal of Contemporary Mathematical Sciences.* 2011. V. 6. P. 585—600.

14. **Tibshirani R.** Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B.* 1994. V. 58. P. 267—288.

15. **Efron B., Hastie T., Tibshirani R.** Least angle regression. *The Annals of Statistics.* 2004. V. 32, N. 2. P. 407—499.

16. **Zou Hui, Hastie T.** Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B.* 2005. V. 67. P. 301—320.

17. **Harrell F. E.** *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* Berlin: Springer, 2001.

18. **Paul R. K.** MulticoUinearity: Causes, Effects and Remedies. — Accessed Apr. 23, 2013, URL: http://bit.lv/lqDHObV.

19. **Strijov V., Krymova E., Weber G.-W.** Evidence optimization for consequently generated models. *Mathematical and Computer Modeling.* 2013. V. 57, N. 1—2. P. 50—56.