Проблема мультиколлинеарности при выборе признаков в регрессионных задачах

А.М. Катруца*, В. В. Стрижов†

Аннотация

В данной работе исследуется проблема мультиколлинеарности и её 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 порождённых 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, на котором основана процедура их тестирования. Также методы выбора признаков сравниваются согласно различным функционалам качества. Проведено сравнение методов выбора признаков в случае наличия в данных определённого типа мультиколлинеарности, и сделан вывод о качестве работы рассматриваемых методов на определённых типах данных.

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

1 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 её обнаружения и устранения описаны в [1, 2, 3]. Также мультиколлинеарность приводит к уменьшению устойчивости оценок вектора параметров. Оценка вектора параметров называется устойчивой,

*Московский физико-технический институт, amkatrutsa@yandex.ru

†Вычислительный центр им. А.А. Дородницына РАН, strijov@gmail.com

если малое изменении 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] применяются методы выбора признаков. Также методы выбора признаков используются 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 [10, 11, 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 используются 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 комбинациями процедур добавления или удаления признаков

Для тестирования методов выбора признаков в [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 даёт 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 количество 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.

2 Постановка задачи выбора признаков

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

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

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-ый компонент вектора регрессионных остатков $\boldsymbol{\varepsilon} = \mathbf{f} - \mathbf{y}$. Обозначим вектор-функцию:

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

Определим функцию ошибки

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

и представление множества индексов элементов выборки в виде:

$$\mathcal{I} = \mathcal{L} \cup \mathcal{C}.$$

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

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

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

Назовём моделью пару $(\mathbf{f}, \mathcal{A})$, где $\mathcal{A} \subset \mathcal{J}$ — 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 $\mathcal{A}$, поэтому вместо $(\mathbf{f}, \mathcal{A})$ для обозначения используемой модели будем использовать $\mathcal{A}$. Таким образом, выбор модели сводится к нахождению оптимального множества индексов $\mathcal{A}^*$ в смысле функции ошибки S, вычисляемой на элементах выборки $\mathfrak{D}_\mathcal{C}$:

$$\mathcal{A}^* = \arg\min_{\mathcal{A}\subset\mathcal{J}} S(\mathcal{A}|\mathbf{w}^*, \mathfrak{D}_\mathcal{C}). \tag{3}$$

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

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}|\mathfrak{D}_\mathcal{L}, \mathcal{A}). \tag{4}$$

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

3 Анализ мультиколлинеарности при выборе признаков

В дальнейшем будем считать, что векторы признаков $\boldsymbol{\chi}_j$ и целевой вектор $\mathbf{y}$ нормированны. Рассмотрим некоторое подмножество $\mathcal{B} \subset \mathcal{J}$ индексов признаков. Назовём признаки *мультиколлинеарными*, если найдутся такие коэффициенты $a_k,\ k \in \mathcal{B}$ и достаточно малое $\delta > 0$, что

$$\left\| \boldsymbol{\chi}_j - \sum_{k\in\mathcal{B}} a_k \boldsymbol{\chi}_k \right\| < \delta, \tag{5}$$

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

Назовём признаки с индексами i, j *коррелирующими*, если найдётся достаточно малое δ_{ij} такое, что:

$$\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\| < \delta_{ij} \tag{6}$$

Из определения следует, что $\delta_{ij} = \delta_{ji}$ и формула (6) есть частный случай формулы (5) при $a_k = 0,\ k \neq j$ и $a_k = 1,\ k = j$.

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

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

3.1 Фактор инфляции дисперсии

Широко известным 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-ым и остальными

признаками. Для нахождения VIF_j необходимо определить оценку $\hat{\mathbf{w}}$ для вектора коэффициентов $\mathbf{w}$ в задаче (1) при $y_i = x_{ij},\ i \in \mathcal{I}$ и $\mathcal{J} = \mathcal{J} \setminus j$. Аналогично (2) определяются RSS и TSS. Величина VIF_j определяется следующим выражением:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

где $R_j^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$ — коэффициент детерминации. Согласно [18] значение $\mathrm{VIF}_j \gtrsim 5$ означает наличие зависимости между j-ым 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 $\mathbf{X}^\intercal\mathbf{X}$, которое равно отношению максимального и минимального по модулю собственных чисел $\lambda_{\max}$ и $\lambda_{\min}$:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Оно показывает насколько матрица $\mathbf{X}^\intercal\mathbf{X}$ близка к вырожденной. Чем больше κ, тем ближе матрица к вырожденной.

3.2 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. Формальной записью предлагаемого критерия является задача (12). 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

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

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

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

Используя сингулярное разложение матрицы $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal,$$

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

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

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

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

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

$$\mathbf{X}^\intercal\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^\intercal\mathbf{U}^\intercal\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^\intercal,$$
$$\mathbf{X}^\intercal\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2.$$

Отношение максимального собственного значение $\lambda_{\max}$ к i-ому собственному значению λ_i назовём 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

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

где $\mathcal{F}_{k-1}$ текущее подмножество признаков.

Оценками дисперсий параметров будут диагональные элементы матрицы $\mathbf{X}^\intercal\mathbf{X}$:

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

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

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

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^*

$$j^* = \arg\max_{j \in \mathcal{F}_{k-1}} q_{i^*j}, \tag{7}$$

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

4 Методы построения тестовых выборок

Для тестирования методов выбора признаков предлагается использовать следующие синтетические выборки. Определим следующие множества задающие структуру выборки:

1) множество ортогональных признаков χ_j с индексами j из множества $\mathcal{P}_f$;

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

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

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

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

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 $j \in \mathcal{C}_f$ мультиколлинеарных между собой признаков, так и из множества $j \in \mathcal{P}_y$ ортогональных целевому вектору $\mathbf{y}$:

$$\langle \mathbf{y}, \boldsymbol{\chi}_j \rangle = 0, \quad j \in \mathcal{J}, \quad \left\| \boldsymbol{\chi}_i - \sum_{l \in \mathcal{B}} \alpha_l \boldsymbol{\chi}_l \right\| < \delta, \quad i \in \mathcal{J}, \quad i \notin \mathcal{B} \subset \mathcal{J}$$
$$\mathcal{J} = \mathcal{P}_y \cap \mathcal{C}_f. \tag{8}$$

   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. Эта случайная величина взята из равномерного распределения на единичном кубе размерности r. Найдётся некоторый признак $\boldsymbol{\chi}_i$ приближающий целевой вектор $\mathbf{y}$:

$$\mathcal{J} = \mathcal{R}, \quad \boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r \sim \mathrm{U}[0,1]^r, \quad \|\mathbf{y} - \boldsymbol{\chi}_i\| < \delta. \tag{9}$$

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

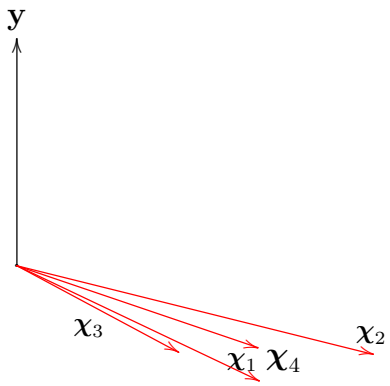


Рис. 1

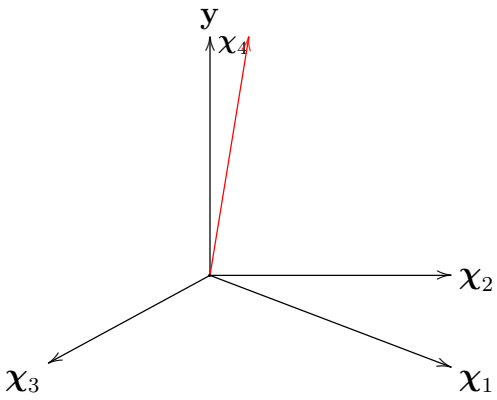


Рис. 2

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

$$\left\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\right\| < \delta_{ij}, \quad i, j \in \mathcal{J}, \quad \|\mathbf{y} - \boldsymbol{\chi}_j\| < \delta, \quad j \in \mathcal{J}, \quad \mathcal{J} = \mathcal{C}_y \tag{10}$$

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

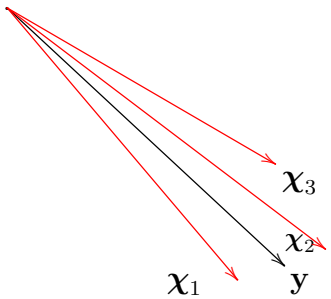


Рис. 3

4. Множество признаков $\boldsymbol{\chi}_j$ с индексами из множества $j \in \mathcal{J}$ состоит из объединения двух множеств: множества ортогональных признаков с индексами из множества $\mathcal{P}_f$

и множества признаков $\boldsymbol{\chi}_c$, коррелированных с некоторыми из них. Индексы c лежат в множестве $\mathcal{C}_f$. При этом целевой вектор $\mathbf{y}$ хорошо приближается линейной комбинацией ортогональных признаков $\boldsymbol{\chi}_j$, $j \in \mathcal{P}_f$:

$$\langle \boldsymbol{\chi}_i, \boldsymbol{\chi}_j \rangle = 0, \quad i, j \in \mathcal{P}_f, \quad \|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\| < \delta_{ij}, \quad i \in \mathcal{P}_f,\ j \in \mathcal{C}_f, \quad \mathbf{y} = \sum_{j \in \mathcal{P}_f} a_j \boldsymbol{\chi}_j, \tag{11}$$

$$\mathcal{J} = \mathcal{P}_f \cup \mathcal{C}_f.$$

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

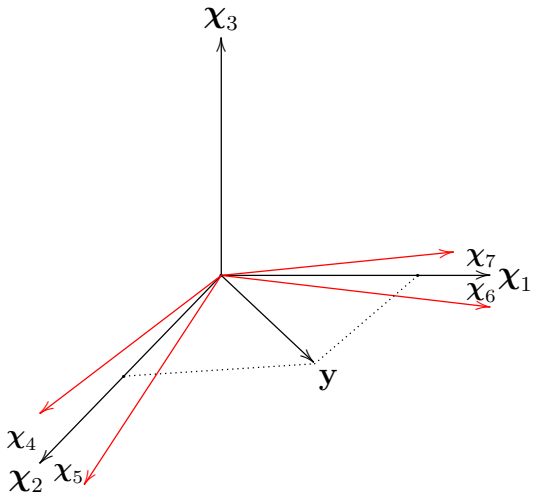


Рис. 4

Комбинируя вышеописанные 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 $\mathcal{P}_f$, $\mathcal{P}_y$, $\mathcal{C}_f$, $\mathcal{C}_y$ и $\mathcal{R}$, можно генерировать выборки для тестирования методов выбора признаков.

5 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 $\mathcal{A} \subset \mathcal{J}$. Для найденного множества признаков получен оптимальный вектор параметров $\mathbf{w}^*_{\mathcal{A}}$. Назовём h максимальную мощность множества индексов признаков $\mathcal{J}_h \subset \mathcal{A}$, при удалении которого значение функции ошибки S не превосходит s_0:

$$h = \underset{S(\mathcal{J}_h, \mathbf{w}_h, \mathfrak{D}) \le s_0}{\arg\max} |\mathcal{J}_h|, \tag{12}$$

где $S(\mathcal{J}_h, \mathbf{w}_h, \mathfrak{D})$ — функция ошибки, в которой первый аргумент — это матрица $\mathbf{X}$ со столбцами, индексы которых лежат в множестве $\mathcal{J}_h$, второй аргумент — вектор параметров $\mathbf{w}_h$, составленный из элементов $\mathbf{w}^*_{\mathcal{A}}$ с индексами из множества $\mathcal{J}_h$ и третий аргумент — выборка. В разделе вычислительный эксперимент определялась величина d равная количеству признаков, удаление которых приводит к ошибке, не превышающей s_0:

$$d = |\mathcal{A}| - h.$$

Определение индексов удаляемых признаков проводилось методом Белсли, задача (7). Методы выбора признаков ранжируются по возрастанию величины d: большие значения d показывают, что выбранное подмножество признаков $\mathcal{A}^*$ (решение задачи (3)) содержит избыточные признаки, удаление которых приводит к росту функции ошибки вплоть до s_0.

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

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

   $$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(m-k)}{\text{TSS}/(m-1)}.$$

   Чем ближе значение R^2_{adj} к единице, тем лучше модель описывает целевой вектор.

2. Критерий C_p позволяет достичь компромисса между величиной RSS и количеством используемых переменных p, а также ликвидировать возможную коллинеарность признаков. Величина C_p определяется следующим образом:

   $$C_p = \frac{\text{RSS}_p}{\text{RSS}} - m + 2p,$$

   где RSS_p — это величина, аналогичная RSS, но найденная при использовании p признаков. Меньшие значения C_p соответствуют лучшему набору признаков.

3. Информационный критерий BIC вычисляется по следующей формуле:

   $$\text{BIC} = \text{RSS} + p \log m.$$

   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

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

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

6 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 $\mathcal{A}^*$, что показывает наличие мультиколлинеарных признаков в множестве отобранных признаков $\mathcal{A}^*$. Эксперименты проводились на выборках при $k = 0.2$ и $k = 0.8$. Для выборок второго типа график зависимости VIF от параметра мультиколлинеарности k и количества избыточных признаков d в множестве отобранных признаков от предельного значения функции ошибки s_0 не строился, так как в этом типе выборок нет мультиколлинеарных признаков.

В экспериментах генерировались выборки четырёх типов, определяемых формулами 8, 9, 10, 11 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 5 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, 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. Величина предельной функции ошибки $s_0 = 0.5$.

Сравнивались 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 параметров $\mathbf{w}^*$. Метод Stepwise последовательно решает задачи (3) и (4), добавляя и удаляя признаки в соответствии с их значимостью, определяемой статистическим тестом. Для алгоритма ElasticNet используется одинаковый вес для штрафа Lasso и Ridge равный 0.5. Прочерк в таблице означает, что метод выбора признаков не отбирает ни один признак и получаемый вектор $\mathbf{w}$ нулевой.

Для выборок первого типа $n = p_y = 50$, $m = 1000$, результаты приведены в таблицах 1 и 2 для $k = 0.2$ и $k = 0.8$ соответственно.

Для выборок второго типа $n = r = 50$, $m = 1000$, результаты приведены в таблице 3. Для выборок третьего типа $n = c_y = 50$, $m = 1000$, результаты приведены в таблицах 4 и 5 для $k = 0.2$ и $k = 0.8$ соответственно.

Для выборок четвёртого типа $p_f = 10$, $c_f = 40$, $m = 1000$, результаты приведены в таблицах 6 и 7 для $k = 0.2$ и $k = 0.8$ соответственно.

Таблица 1 Значения функционалов качества для выборок первого типа при $k = 0.2$.

	d	C_p	RSS	κ	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	0	-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 $k = 0.8$

	d	C_p	RSS	κ	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

На рисунках 5, 6, 7 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 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 наблюдается резкое уменьшение величины VIF. Это говорит об отсутствии линейной зависимости между выбранными признаками в выборках, сгенерированных при $k \gtrsim 0.4$.

На рис. 7 показана зависимость VIF от параметра мультиколлинеарности k для четвёртого типа выборок при работе различных методов. Метод LARS показывает резкие скачки значений VIF (рис. 7 а)). Это не позволяет оценить зависимость VIF от k для других методов, поэтому на рис. 7 б) изображены аналогичные графики для всех рассматриваемых методов, кроме LARS. Методы Lasso и ElasticNet демонстрируют скачки, схожие со скачками у LARS, но меньшей амплитуды. Поэтому зависимости VIF от k, получаемые после использования методов Stepwise и Ridge, изображены на рис. 7 в) и 7 г). Для выборок четвёртого типа после применения метода Stepwise значения VIF не превышают двух при росте коэффициента k. Это означает, что метод Stepwise для выборок четвёртого типа

Таблица 3 Значения функционалов качества для выборок второго типа

	d	C_p	RSS	κ	VIF	R^2_{adj}	BIC	p-value
Lasso	0	$7 \cdot 10^6$	$8.50 \cdot 10^{-4}$	1	0.25	1	6.9	0
Elastic Net	0	$8.76 \cdot 10^{-4}$	$8.76 \cdot 10^{-4}$	1	0.25	1	6.9	0
Ridge	0	$7.97 \cdot 10^9$	0.97	1	0.25	-3	7.88	0
LARS	0.2	-997	$1.3 \cdot 10^{-10}$	2.19	0.32	1	8.29	0
Stepwise	4.6	-997	$1.33 \cdot 10^{-10}$	28.86	0.89	1	53.88	0

Таблица 4 Значения функционалов качества для выборок третьего типа при $k = 0.2$

	d	C_p	RSS	κ, $\cdot 10^8$	VIF, $\cdot 10^7$	R^2_{adj}	BIC	p-value
Ridge	0	$2.3 \cdot 10^9$	0.97	24	1.14	-3.17	346.36	0
Lasso	1	$2 \cdot 10^6$	$8.5 \cdot 10^{-4}$	0.95	0.58	1	13.82	0
Elastic Net	3.2	$2 \cdot 10^6$	$8.5 \cdot 10^{-4}$	2.8	0.97	1	41.45	0
Stepwise	36	-997	$4.22 \cdot 10^{-10}$	24	1.14	1	345.39	0
LARS	36	-997	$4.22 \cdot 10^{-10}$	24	1.14	1	345.39	0

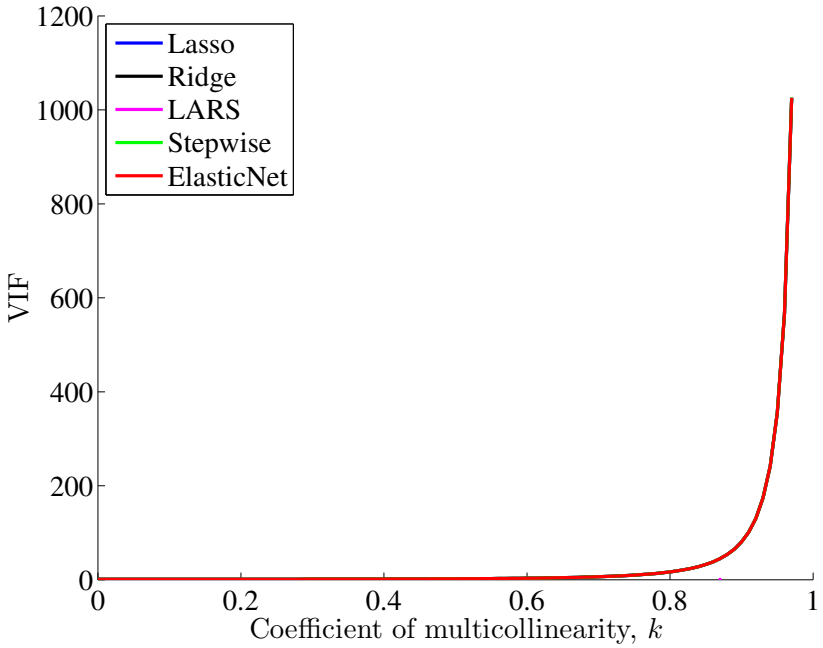


Рис. 5 Зависимость величины фактора инфляции дисперсии VIF от параметра мультиколлинеарности k для первого типа выборок

даёт набор признаков, в котором нет линейной зависимости между признаками.

Рассмотрим зависимость количества мультиколлинеарных признаков d в множестве отобранных признаков от значений предельной ошибки s_0 для ранее рассмотренных типов выборок на рис. 8, 9, 10.

На рис. 8 показана зависимость количества лишних признаков d в множестве отобран-

Таблица 5 Значения функционалов качества для выборок третьего типа при $k = 0.8$

	d	C_p	RSS	κ	VIF	R^2_{adj}	BIC	p-value
Lasso	0	$5.16 \cdot 10^8$	$8.5 \cdot 10^{-4}$	1	0.24	1	6.9	0
Ridge	0	$5.9 \cdot 10^{11}$	0.97	$6.07 \cdot 10^{11}$	$2.9 \cdot 10^9$	-3.17	346.36	0
Elastic Net	3.2	$5.16 \cdot 10^8$	$8.5 \cdot 10^{-4}$	$7.3 \cdot 10^{10}$	$2.5 \cdot 10^9$	1	41.45	0
Stepwise	36	-997	$1.73 \cdot 10^{-12}$	$6.07 \cdot 10^{11}$	$2.9 \cdot 10^9$	1	345.39	0
LARS	36	-997	$1.65 \cdot 10^{-12}$	$6.07 \cdot 10^1 1$	$2.9 \cdot 10^9$	1	345.39	0

Таблица 6 Значения функционалов качества для выборок четвёртого типа при $k = 0.2$

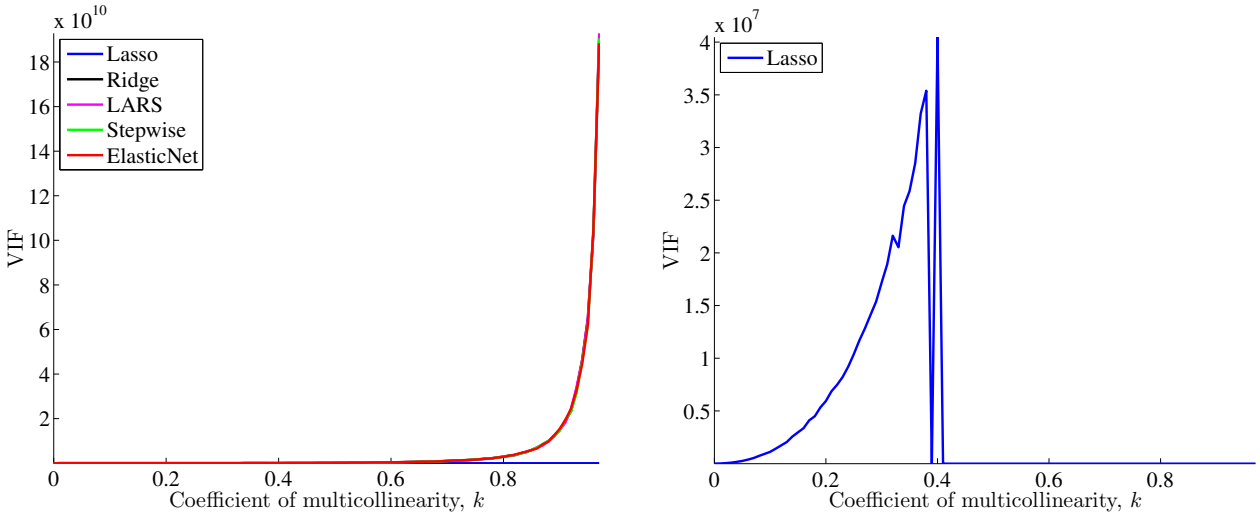
	d	C_p	RSS	κ	VIF	R^2_{adj}	BIC	p-value
Ridge	0	$6 \cdot 10^{30}$	0.95	$8.42 \cdot 10^{15}$	$1.15 \cdot 10^{23}$	-3	210.95	0
Stepwise	1	-868.95	$5.45 \cdot 10^{-29}$	1	0.63	1	13.82	0
LARS	1.8	$5.38 \cdot 10^{29}$	0.38	$2.1 \cdot 10^{16}$	$3.3 \cdot 10^{30}$	-0.62	102.62	0
Lasso	18	$5.84 \cdot 10^{27}$	$9.18 \cdot 10^{-4}$	$1.4 \cdot 10^{16}$	$5.32 \cdot 10^{20}$	1	150.6	0
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	0

Таблица 7 Значения функционалов качества для выборок четвёртого типа при $k = 0.8$

	d	C_p	RSS	κ	VIF	R^2_{adj}	BIC	p-value
Ridge	0	$1.8 \cdot 10^{30}$	0.95	10^{16}	$8.65 \cdot 10^{16}$	-2.97	152.92	0
Stepwise	1	$9.4 \cdot 10^5$	$8.8 \cdot 10^{-25}$	1	0.63	1	13.82	0
LARS	1.2	10^{30}	0.38	$3 \cdot 10^{29}$	10^{20}	-0.57	108.15	0
Lasso	14.8	$1.73 \cdot 10^{27}$	$9.2 \cdot 10^{-4}$	$9.92 \cdot 10^{15}$	10^{17}	1	150.59	0
Elastic Net	15.2	$1.7 \cdot 10^{27}$	$9.2 \cdot 10^{-4}$	$9.92 \cdot 10^{15}$	10^{17}	1	150.59	0

ных признаков от предельного значения функции ошибки s_0 для первого типа выборок при значениях $k = 0.2$ и $k = 0.8$. Величина d стабильно равна нулю из-за ортогональности целевого вектора и всех признаков вплоть до значений s_0 близкими к единице. Далее идёт резкий скачок d, так как предельное значение функции ошибки выросло достаточно, чтобы удалить сразу почти все признаки.

На рис. 9 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 или два признака, наилучшим образом приближающие целевой вектор, поэтому величина d для этого метода равна нулю или единице. Аналогично, но чуть хуже работает метод ElasticNet, 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 s_0 достаточна велика, чтобы удалять хоть одни признак, но как только s_0 приближается к единице становится возможным удалить сразу почти все признаки. Для методов LARS и Sterwise наблюда-



а) б)

Рис. 6 Зависимость фактора инфляции дисперсии VIF от параметра мультиколлинеарности k для третьего типа выборок при работе: а) всех рассматриваемых методов отбора признаков, б) только Lasso

ется 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. 10 показана зависимость d от параметра s_0 для четвёртого типа выборок при $k = 0.2$ и $k = 0.8$. Наиболее стабильные решения даёт метод Stepwise, у которого в среднем обнаруживается только один признак, удаление которого приводит к ошибке, не превышающей s_0. Чуть хуже работает метод LARS: количество 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 единице, а затем стабилизация на уровне $d \simeq 20$. Для метода Ridge вид зависимости схож с предыдущими типами выборок, только для четвёртого типа после преодоления s_0 значения равного единице величина d начала сильно колебаться. Это показывает неустойчивость набора признаков, получаемого методом Ridge для четвёртого типа выборок.

7 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 четвёртого 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

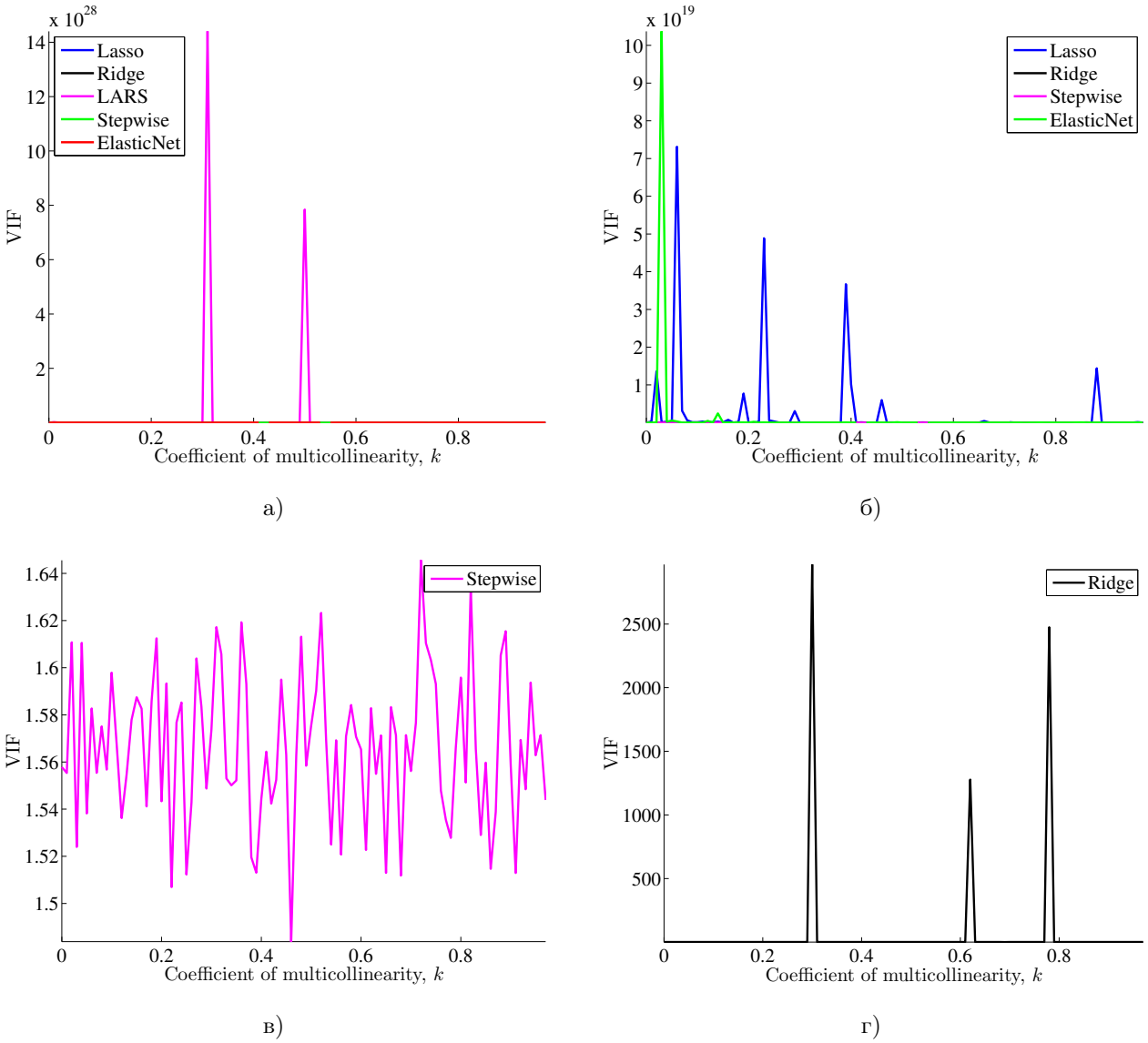
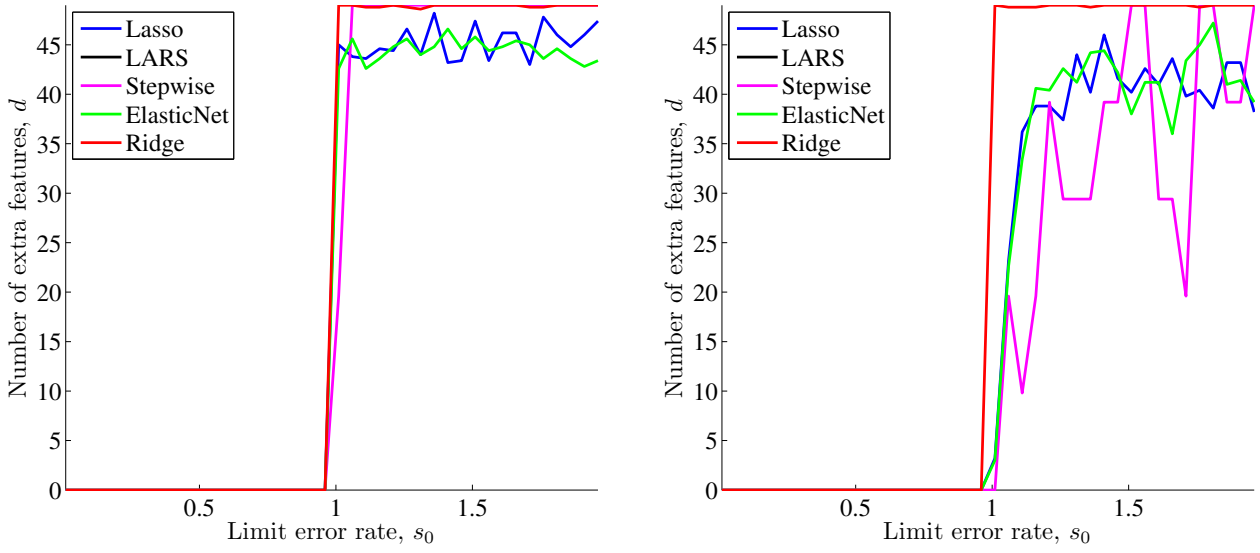


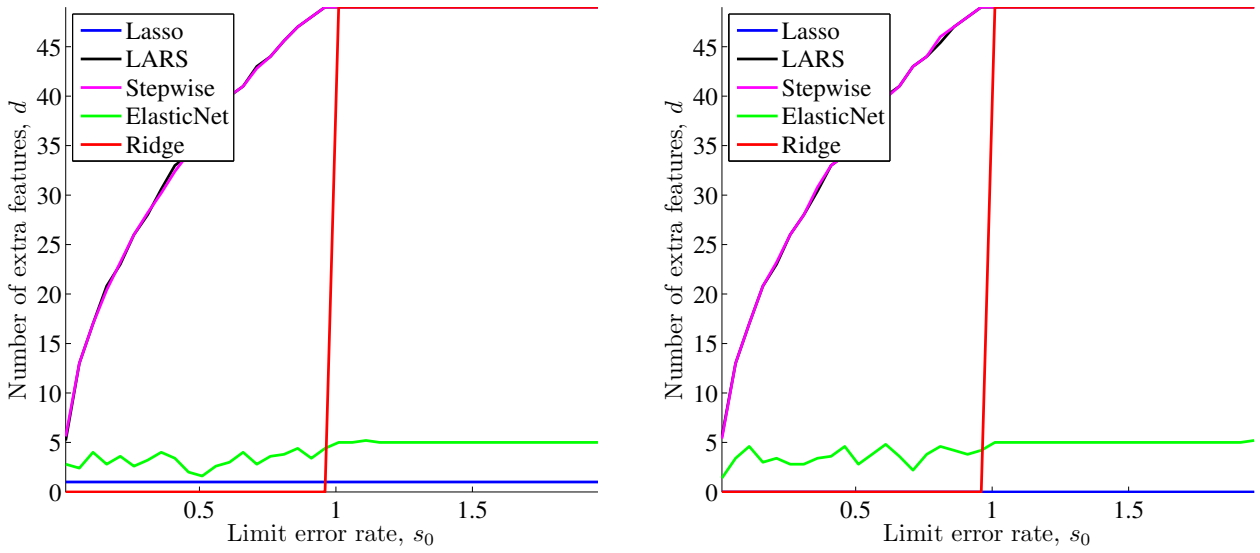
Рис. 7 Зависимость величины фактора инфляции дисперсии VIF от параметра мультиколлинеарности k для четвёртого типа выборок при работе: а) всех методов отбора признаков, б) всех методов кроме LARS, в) метода Stepwise, г) метода Ridge

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 даёт метод Lasso, для выборок четвёртого типа — методы LARS и Stepwise.



а) б)

Рис. 8 Зависимость количества мультиколлинеарных признаков d в множестве отобранных признаков от предельного значения функции ошибки s_0 для первого типа выборок при: а) $k = 0.2$, б) $k = 0.8$



а) б)

Рис. 9 Зависимость количества мультиколлинеарных признаков d в множестве отобранных признаков от предельного значения функции ошибки s_0 для третьего типа выборок при: а) $k = 0.2$, б) $k = 0.8$

Список литературы

[1] R. G. Askin. Multicollinearity in regression: Review and examples. *Journal of Forecasting*, 1(3):281—292, 1982.

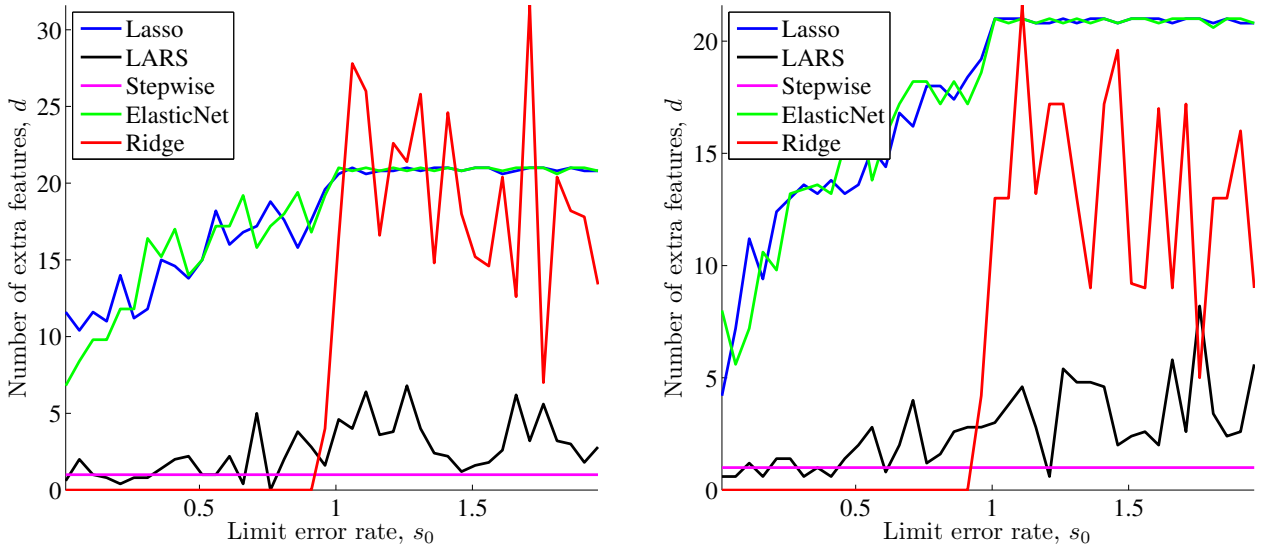


Рис. 10 Зависимость количества мультиколлинеарных признаков d в множестве отобранных признаков от предельного значения функции ошибки s_0 для четвёртого типа выборок при: а) $k = 0.2$, б) $k = 0.8$

[2] Edward E Leamer. Multicollinearity: A bayesian interpretation. *The Review of Economics and Statistics*, 55(3):371–80, 1973.

[3] David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, New York, 2005.

[4] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, Washington D.C., 2003.

[5] В. В. Стрижов, М. П. Кузнецов, and К. В. Рудаков. Метрическая кластеризация последовательностей аминокислотных остатков в ранговых шкалах. *Математическая биология и биоинформатика*, 7(1):345–359, 2012.

[6] Yi-Wei Chen and Chih-Jen Lin. Combining svms with various feature selection strategies. In *Feature Extraction*, pages 315–324. Springer, 2006.

[7] George H John, Ron Kohavi, Karl Pfleger, et al. Irrelevant features and the subset selection problem. In *ICML*, volume 94, pages 121–129, 1994.

[8] Vorontsov K. Combinatorial probability and the tightness of generalization bounds. *Pattern Recognition and Image Analysis*, 18(2):243—259, 2008.

[9] Il-Gyo Chong and Chi-Hyuck Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1–2):103 – 112, 2005.

[10] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157—1182, 2003.

[11] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.*, 34(3):483–519, 2013.

[12] L Ladha and T Deepa. Feature selection methods and algorithms. *International Journal on Computer Science & Engineering*, 3(5), 2011.

[13] M. El-Dereny and N. I. Rashwan. Solving multicollinearity problem using ridge regression models. *Int. Journal of Contemp. Math. Sciences*, 6:585 — 600, 2011.

[14] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[15] B. Efron, T. Hastie, and R. Tibshirani. Least angle regression. *Ann. Statist*, pages 407–499, 2004.

[16] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

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

[18] Ranjit Kumar Paul. Multicollinearity: Causes, effects and remedies. Technical report, Working paper, unknown date. Accessed Apr. 23, 2013, http://pb8.ru/7hy, 2006.

[19] Strijov Vadim, Krymova Ekaterina, and Weber Gerhard-Wilhelm. Evidence optimization for consequently generated models. *Mathematical and Computer Modelling*, 57(1-2):50–56, 2013.