

# Анализ метода отбора признаков QPFS для обобщенно-линейных моделей\*

А. Д. Толмачев, А. А. Адуенко, В. В. Стрижов

tolmachev.a.d.@phystech.edu; aduenko1@gmail.com; strijov@ccas.ru

В данной работе исследуется проблема мультиколлинеарности и её влияние на точность методов выбора признаков. Решается задача выбора признаков и различные подходы к ее решению. Проведен анализ метода отбора признаков на основе квадратичного программирования. В работе приводятся критерии сравнения различных способов отбора признаков, и проведено сравнение различных методов на тестовых выборках. Сделан вывод об эффективности рассматриваемых подходов на синтетических данных.

**Ключевые слова:** регрессионный анализ; мультиколлинеарность; число обусловленности; выбор признаков; квадратичное программирование

## 1 Введение

Работа посвящена анализу метода отбора признаков QPFS на основе квадратичного программирования и сравнительному анализу различных методов отбора 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. Это явление затрудняет оценивание регрессионных параметров и выявление зависимости между признаками и целевым вектором. Проблема мультиколлинеарности и возможные способы её обнаружения и устранения описаны в [1, 10, 14].

Задача выбора оптимального подмножества признаков является одной из основных частей выбора модели в исследуемом методе обучения (см. в [3]). Методы выбора признаков основаны на минимизации некоторого функционала, который отражает качество рассматриваемого подмножества признаков. В [6, 7] сделан обзор основных существующих методов отбора признаков.

В [9, 13] предложен новый метод отбора признаков, использующий один из основных методов оптимизации, квадратичное программирование (см. [13]), для задачи отбора признаков. Цель данной работы состоит в анализе возможностей применения метода квадратичного программирования в задаче выбора признаков.

Важной частью этой работы является сравнение различных методов отбора признаков, описанных, например, в [8], на различных тестовых выборках.

## 2 Постановка задачи

### 2.1 Рассматриваемая модель

Задана выборка  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ , где  $i \in \{1, 2, \dots, m\}$ , где множество свободных переменных — вектор  $\mathbf{x} = [x_1, x_2, \dots, x_j, \dots, x_n]$ , где  $j \in \mathcal{J} = \{1, 2, \dots, n\}$ . Предполагается, что  $\mathbf{x}_i \in \mathbb{X} \subset \mathbb{R}^n$  и  $y_i \in \mathbb{Y} \subset \mathbb{R}$ .

Введем обозначения  $\mathbf{y} = [y_1, y_2, \dots, y_m]^T$  — вектор значений зависимой переменной (целевой вектор),  $\chi_j = [x_{1j}, x_{2j}, \dots, x_{mj}]^T$  — реализация  $j$ -ой свободной переменной ( $j$ -ый признак), и  $\mathbf{X} = [x_1^T, x_2^T, \dots, x_n^T]^T = [\chi_1, \chi_2, \dots, \chi_n]$  — матрица плана эксперимента.

\*Работа выполнена в рамках курса «Моя первая научная статья», НИУ МФТИ, 2021

Предполагается, что вектор  $\mathbf{x}_i$  и значение целевой переменной  $y_i$  связаны соотношением

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

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

$$\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{X}) = [f(\mathbf{w}, \mathbf{x}_1), f(\mathbf{w}, \mathbf{x}_2), \dots, f(\mathbf{w}, \mathbf{x}_m)]^T \in \mathbb{Y}^m.$$

Назовем моделью пару  $(\mathbf{f}, \mathcal{A})$ , где  $\mathcal{A} \subset \mathcal{J}$  — подмножество индексов признаков, используемое для вычисления вектор-функции  $\mathbf{f}$ . Предполагается гомоскедастичность модели, т.е.  $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$ .

## 2.2 Применение квадратичной оптимизации для задачи отбора признаков

В [8, 13] предлагается подход с применением квадратичной оптимизации для задачи выбора признаков в сформулированной выше модели. Основная идея предлагаемого подхода заключается в минимизации количества схожих признаков и максимизации количества релевантных признаков. Пусть  $\mathcal{J}$  — множество признаков в рассматриваемой модели, и  $|\mathcal{J}| = n$ .

Положим,  $\mathbf{Q} \in \mathbb{R}^{n \times n}$  — матрица попарных корреляций Пирсона между признаками  $[\chi_1, \chi_2, \dots, \chi_n]$ , а  $\mathbf{b} \in \mathbb{R}^n$  — вектор корреляций Пирсона между признаками  $[\chi_1, \chi_2, \dots, \chi_n]$  и целевым вектором  $\mathbf{y} = [y_1, y_2, \dots, y_m]^T$ .

В [8] рассматривается функционал  $Q(\mathbf{a}) = \mathbf{a}^T \mathbf{Q} \mathbf{a} - \mathbf{b}^T \mathbf{a}$ , где  $\mathbf{a} \in \mathbb{R}^n$  и  $\mathbf{Q} \in \mathbb{R}^{n \times n}$  — матрица схожести признаков, а  $\mathbf{b} \in \mathbb{R}^n$  — вектор релевантности признаков, определенные выше. Матрицу  $\mathbf{Q}$  и вектор  $\mathbf{b}$  будем представлять как функции Sim и Rel соответственно, где Sim:  $\mathcal{J} \times \mathcal{J} \rightarrow [0, 1]$ , Rel:  $\mathcal{J} \rightarrow [0, 1]$ . Таким образом, необходимо решить задачу оптимизации:

$$\mathbf{a}^* = \arg \min_{\mathbf{a} \in \mathbb{B}^n} Q(\mathbf{a}).$$

Важно отметить, что задача целочисленного квадратичного программирования, сформулированная выше, является **NP**-полной, так как поиск минимума функции  $\mathbf{Q}$  ведется по вершинам булева куба  $\mathbb{B}^n = \{0, 1\}^n$ . Поэтому, чтобы можно было применять различные методы выпуклой оптимизации будем искать минимум функции по выпуклой оболочке булева куба  $\text{Conv}(\mathbb{B}^n) = [0, 1]^n$ .

Тогда получаем следующую задачу выпуклой оптимизации:

$$\begin{cases} \mathbf{z}^* = \arg \min_{\mathbf{z} \in [0, 1]^n} \mathbf{z}^T \mathbf{Q} \mathbf{z} - \mathbf{b}^T \mathbf{z} \\ \|\mathbf{z}\|_1 \leq 1 \end{cases}$$

В рассматриваемом в [8] методе не используется коэффициент для балансировки между частями минимизируемого функционала, что вносит неопределенность в этот метод. Поэтому далее будем рассматривать оптимизационную задачу, предложенную в [13], где добавлена балансировка:

$$\begin{aligned} & \frac{1}{2}(1-\alpha)\mathbf{a}^\top \mathbf{Q}\mathbf{a} - \alpha\mathbf{b}^\top \mathbf{a} \rightarrow \min_{\mathbf{a}} \\ & \text{s.t. } \mathbf{a} \geq 0, \sum_{i=1}^n a_i = 1. \end{aligned} \quad (1)$$

Решение задачи (1)  $\mathbf{a}^*$  определяет, какие признаки используются при построении модели. Признак  $j$  активен  $\iff a_j > 0$ .

Эта задача эквивалентна

$$\begin{aligned} & \frac{1}{2}\mathbf{a}^\top \mathbf{Q}\mathbf{a} - \frac{\alpha}{1-\alpha}\mathbf{b}^\top \mathbf{a} \rightarrow \min_{\mathbf{a}} \\ & \text{s.t. } \mathbf{a} \geq 0, \sum_{i=1}^n a_i = 1. \end{aligned} \quad (2)$$

Обозначим задачу (2) как  $S\left(\underbrace{\frac{\alpha}{1-\alpha}}_{\beta^{-1}}, 1\right)$ , где 1 указывает на норму решения. Далее

рассмотрим задачу  $S(\beta^{-1}, \gamma)$  и сделаем замену переменной  $\mathbf{a} = \gamma\tilde{\mathbf{a}}$ , получим

$$\begin{aligned} & \gamma^2 \left( \frac{1}{2}\tilde{\mathbf{a}}^\top \mathbf{Q}\tilde{\mathbf{a}} - \frac{1}{\beta\gamma}\mathbf{b}^\top \tilde{\mathbf{a}} \right) \rightarrow \min_{\tilde{\mathbf{a}}} \\ & \text{s.t. } \tilde{\mathbf{a}} \geq 0, \sum_{i=1}^n \tilde{a}_i = 1, \end{aligned} \quad (3)$$

откуда задача  $S(\beta^{-1}, \gamma)$  эквивалентна задаче  $S((\beta\gamma)^{-1}, 1)$  в терминах активных признаков (сами же решения отличаются в  $\gamma$  раз), а потому мы имеем дело именно с однопараметрическим семейством и задание нормы решения, равной одному, не ограничивает общности.

Рассмотрим теперь замену переменной  $\mathbf{a} = \beta^{-1}\tilde{\mathbf{a}}$  в (2). Получим, что задача (2) эквивалентна при  $\alpha \in (0, 1)$  задаче

$$\begin{aligned} & \frac{1}{2}\tilde{\mathbf{a}}^\top \mathbf{Q}\tilde{\mathbf{a}} - \mathbf{b}^\top \tilde{\mathbf{a}} \rightarrow \min_{\tilde{\mathbf{a}}} \\ & \text{s.t. } \tilde{\mathbf{a}} \geq 0, \sum_{i=1}^n \tilde{a}_i = \beta. \end{aligned} \quad (4)$$

Наряду с задачей (4) можно рассмотреть задачу

$$\begin{aligned} & \frac{1}{2}\tilde{\mathbf{a}}^\top \mathbf{Q}\tilde{\mathbf{a}} - \mathbf{b}^\top \tilde{\mathbf{a}} \rightarrow \min_{\tilde{\mathbf{a}}} \\ & \text{s.t. } \tilde{\mathbf{a}} \geq 0, \sum_{i=1}^n \tilde{a}_i \leq \beta \end{aligned} \quad (5)$$

и соответствующую задачу без ограничения на норму вектора  $\tilde{\mathbf{a}}$ :

$$\begin{aligned} & \frac{1}{2}\tilde{\mathbf{a}}^\top \mathbf{Q}\tilde{\mathbf{a}} - \mathbf{b}^\top \tilde{\mathbf{a}} \rightarrow \min_{\tilde{\mathbf{a}}} \\ & \text{s.t. } \tilde{\mathbf{a}} \geq 0. \end{aligned} \quad (6)$$

### 3 Теоретическая часть

#### 3.1 Сравнение различных задач

Задача (1) не позволяет выбросить все признаки, поскольку  $\|\mathbf{a}\| = 1 > 0$ . Задача (5) эквивалентна ограничению неравенства в исходной задаче и, например, при  $\alpha = 0$  будет иметь решением исключение всех признаков. Далее приведем анализ свойств решения (1) с ограничением равенства и неравенства для разных значений  $\alpha$ .

**Таблица 1** Свойства решения в методе QPFS в зависимости от параметра  $\alpha$

$\alpha$	Равенство	Неравенство
$\alpha = 0$	Используются все признаки, $\mathbf{Qa}^* = \eta \mathbf{e}$ , $\eta > 0$	Выброшены все признаки
$\alpha \rightarrow 0$	Используются все признаки, $\mathbf{Qa}^* \rightarrow \eta \mathbf{e}$ , $\eta > 0$	Решение (6)
$\alpha \rightarrow 1$	Сходимся к отбору одного признака с максимальным $b_j$	То же, что и в «равенство»
$\alpha = 1$	Отбор одного признака с максимальным $b_j$	То же, что и в «равенство»

В случае  $\alpha = 0$  для неравенства  $(\sum_{i=1}^n a_i \leq 1)$  решением является  $\mathbf{a}^* = 0$ . В случае равенства  $(\sum_{i=1}^n a_i = 1)$ , чтобы минимизировать потери от необходимости иметь ненулевой  $\mathbf{a}$ , в оптимальном  $\mathbf{a}^*$  оптимизируемая функция  $\frac{1}{2} \mathbf{a}^T \mathbf{Qa}$  должна иметь одинаковый градиент по всем направлениям (так как иначе можно уменьшить одну координату, увеличить другую, оставив норму  $\mathbf{a}$  неизменной, уменьшив значение функции). Тот же результат можно получить и другим способом - из рассмотрения Лагранжиана задачи.

#### 3.2 Анализ решения QPFS в зависимости от параметра $\alpha$

Задачи (4) и (5) эквивалентны для некоторых  $\eta$  задаче

$$\frac{1}{2} \tilde{\mathbf{a}}^T \mathbf{Q} \tilde{\mathbf{a}} - \mathbf{b}^T \tilde{\mathbf{a}} + \eta \sum_{j=1}^n \mathbf{a}_j \rightarrow \min_{\tilde{\mathbf{a}}} \text{ s.t. } \tilde{\mathbf{a}} \geq 0,$$

что эквивалентно

$$\frac{1}{2} \tilde{\mathbf{a}}^T \mathbf{Q} \tilde{\mathbf{a}} - \mathbf{b}^T \tilde{\mathbf{a}} \rightarrow \min_{\tilde{\mathbf{a}}} \text{ s.t. } \tilde{\mathbf{a}} \geq 0,$$

где  $\tilde{b}_j = b_j - \eta$  и  $\eta$  монотонно убывает по  $\|\tilde{\mathbf{a}}\|_1 = \beta$ . При этом для случая неравенства  $\eta \geq 0$ , в для равенства  $\eta < 0$ , если  $\beta > \|\mathbf{a}^*\|_1$ , где  $\mathbf{a}^*$  есть решение задачи (6) без ограничения на норму.

Таким образом, добавление ограничения на норму, фактически штрафует релевантность и происходит исключение тех признаков, у которых  $b_j < \eta$ , поскольку у них поправленная релевантность  $\tilde{b}_j$  становится отрицательной.

#### 3.3 Связь с Lasso-моделью для линейной регрессии

Стандартная задача линейной регрессии с  $L_1$ -регуляризацией (Lasso) имеет вид

$$\frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \tau \|\mathbf{w}\|_1 \rightarrow \min_{\mathbf{w}}. \quad (7)$$

Предположим, что  $\mathbf{x}_j^T \mathbf{x}_j = 1$ ,  $\mathbf{y}^T \mathbf{y} = 1$ , то есть признаки и целевая переменная нормированы. В качестве функций сходства (Sym) и релевантности (Rel) в QPFS рассмотрим

корреляцию Пирсона, как сказано в постановке задачи. Имеем

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \tau\|\mathbf{w}\|_1 = \frac{1}{2}\underbrace{\mathbf{y}^\top \mathbf{y}}_{=1} + \frac{1}{2}\mathbf{w}^\top \underbrace{\mathbf{X}^\top \mathbf{X}}_{\tilde{\mathbf{Q}}} \mathbf{w} - \underbrace{(\mathbf{X}^\top \mathbf{y})^\top}_{\tilde{\mathbf{b}}} \mathbf{w},$$

где учтена нормированность всех признаков и целевой переменной, поэтому корреляция и ковариация совпадают; Здесь,  $\tilde{\mathbf{Q}}$ ,  $\tilde{\mathbf{b}}$  – корреляции между признаками и признаками и целевой переменной соответственно (см. раздел 2.2). Отсюда задачу (7) можно переписать в виде

$$\frac{1}{2}\mathbf{w}^\top \tilde{\mathbf{Q}}\mathbf{w} - \tilde{\mathbf{b}}^\top \mathbf{w} + \tau\|\mathbf{w}\|_1 \rightarrow \min_{\mathbf{w}},$$

что может быть переписано в эквивалентную задачу с ограничением равенства (так же для неравенства) для некоторого  $\eta$ :

$$\begin{aligned} \frac{1}{2}\mathbf{w}^\top \tilde{\mathbf{Q}}\mathbf{w} - \tilde{\mathbf{b}}^\top \mathbf{w} &\rightarrow \min_{\mathbf{w}} \\ \text{s.t. } \|\mathbf{w}\|_1 &= \eta. \end{aligned} \tag{8}$$

Если все корреляции Пирсона между признаками и между признаками и целевой переменной неотрицательны (то есть векторы  $\mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{y}$  лежат в одном многомерном квадранте), то  $\tilde{\mathbf{Q}} = \mathbf{Q}$ ,  $\tilde{\mathbf{b}} = \mathbf{b}$ , то есть задача (8) тождественна QPFS, но без ограничения  $\mathbf{w} \geq 0$ . Таким образом, QPFS можно рассматривать как Lasso ограничением на неотрицательность весов, если все корреляции Пирсона между признаками и между признаками и целевой переменной неотрицательны. Далее, если истинный вектор весов в линейной регрессии  $\mathbf{w} \geq 0$ , то условие на неотрицательность оценки весов тоже избыточно (так как  $\mathbf{w}^*$  в задаче (8) и так будет неотрицательным, начиная с некоторого размера выборки), и QPFS будет полностью тождественен Lasso-модели для линейной регрессии. Значит, получаем условия тождественности QPFS( $\alpha$ ) методу lasso( $\tau$ ), где между  $\alpha$  и  $\tau$  существует некоторая связь:

- Rel = Sim = |Pearson correlation|;
- Нормированность признаков и целевой переменной:  $\mathbf{y}^\top \mathbf{y} = \mathbf{x}_j^\top \mathbf{x}_j = 1$ ;  $j = 1, \dots, n$ ;
- Неотрицательность попарных корреляций:  $\mathbf{y}^\top \mathbf{x}_j \geq 0$ ,  $\mathbf{x}_j^\top \mathbf{x}_l \geq 0$ ;  $j, l = 1, \dots, n$ ;
- Неотрицательность истинного вектора весов  $\mathbf{w}^* \geq 0$ .

Отметим, что в Lasso-модели реализуется ситуация, в которой исключаются все признаки, где  $\tau \rightarrow \infty$ , что соответствует QPFS с ограничением неравенства при  $\alpha = 0$ . Кроме того, “закритический” режим из QPFS с равенством, когда  $\|\mathbf{a}\|_1 = \beta > \|\mathbf{a}^*\|$ , где  $\mathbf{a}^*$  есть решение (6), в lasso не реализуется, поскольку, штраф за норму приводит всегда к ее сокращению, а потому  $\tau \rightarrow 0$  в lasso соответствует  $\beta \rightarrow \|\mathbf{a}^*\|$  в QPFS с ограничением типа равенства.

## 4 Базовый эксперимент

### 4.1 Проблемы метода QPFS

Благодаря условию на неотрицательность коэффициентов  $\mathbf{a} \geq 0$  в задаче QPFS (4), штраф на  $\|\mathbf{a}\|_1$  становится штрафом на сумму коэффициентов, что упрощает оптимизацию. Авторы оригинального метода [13] прямо указывают на скорость оптимизации как на основное преимущество метода QPFS при сопоставимом качестве прогноза (например, с Lasso-моделью) на тестовых выборках в рассмотренных наборах данных.

Как показано в предыдущей главе, при выполнении некоторых условий (в частности, неотрицательности истинного вектора параметров модели  $\mathbf{w}^*$ ), QPFS будет в точности эквивалентен методу Lasso, то есть метод обладает лучшими свойствами с точки зрения оптимизации, при этом давая то же решение, что и Lasso. Однако, когда условия эквивалентности не выполнены, качество метода становится хуже, чем Lasso, поскольку не учитывает, например, что релевантность пары признаков может быть значительно выше, чем релевантность каждого из них.

## 4.2 Первый эксперимент

Рассмотрим выборку  $(\mathbf{X}, \mathbf{y})$  в признаковом пространстве размерности  $n = 2$ . В исследуемой модели будут два признака  $x_1, x_2$  и целевая переменная  $y$ . Пусть  $x_1 \sim \mathcal{N}(0, 1)$ ,  $y \sim \mathcal{N}(0, 1)$ , а  $x_2 = x_1 + \varepsilon \cdot y$ , т.е.  $x_1$  и  $y$  - независимые случайные величины из стандартного нормального распределения, а  $\varepsilon$  - заранее выбранное малое значение, где  $\varepsilon > 0$ . Таким образом, мы получаем, что  $y_i = \frac{x_{2i} - x_{1i}}{\varepsilon}$ .

Истинная корреляция Пирсона первого признака и целевой переменной  $y$  равна 0, а корреляция со вторым - равна  $\varepsilon/\sqrt{1 + \varepsilon^2}$ , схожесть двух признаков  $1/\sqrt{1 + \varepsilon^2}$ . При малом  $\varepsilon$  выборочная корреляция обоих признаков с целевой переменной будет мала, а схожесть двух признаков - велика. При этом надежное восстановление целевой переменной  $y$  возможно только при наличии обоих признаков в выборке, что соответствует ситуации с отсутствием отбора признаков (малое  $\alpha$ ).

Добавим теперь в выборку  $N$  шумовых признаков. Истинное сходство каждого из таких признаков с целевой переменной равно 0 (такое же как и для признака 1) и с учетом того, что QPFS не учитывает взаимодействия между признаками, признаки 1 и 2 не имеют значительного преимущества по отношению к шумовым в терминах релевантности целевой переменной (признак 1 в точности шумовой в изоляции, так как независим от  $y$ ). При этом признаки 1 и 2 получают штраф за "похожесть" друг на друга. По этой причине при работе QPFS либо происходит исключение одного или обоих признаков 1 и 2 при исключении некоторых или всех шумовых, или оба признака 1 и 2 остаются, но вместе с ними остаются почти все или все шумы (см. рис. 2). В то же время Lasso учитывает взаимосвязи между признаками и не требует неотрицательности коэффициентов (ссылка на сравнение на этом датасете QPFS и Lasso). В рассматриваемом примере не выполнено одно из условий эквивалентности Lasso и QPFS, т.к. среди компонент вектора  $\mathbf{w}^* = (-1/\varepsilon, 1/\varepsilon)^T$  есть отрицательные значения, т.к. одно из чисел  $-1/\varepsilon, 1/\varepsilon$  меньше нуля.

На рис. 2 приведен график отбора признаков методом QPFS в данном примере при  $N = 10$ . При фиксированном значении порога  $\tau \in [0, 1]$  было найдено количество основных (а их всего 2:  $\mathbf{x}_1, \mathbf{x}_2$ ) и шумовых признаков, при которых  $a_j > \tau$ . Видим, что если отбираются оба основных признака, то вместе с ними отбирается и много шумовых признаков.

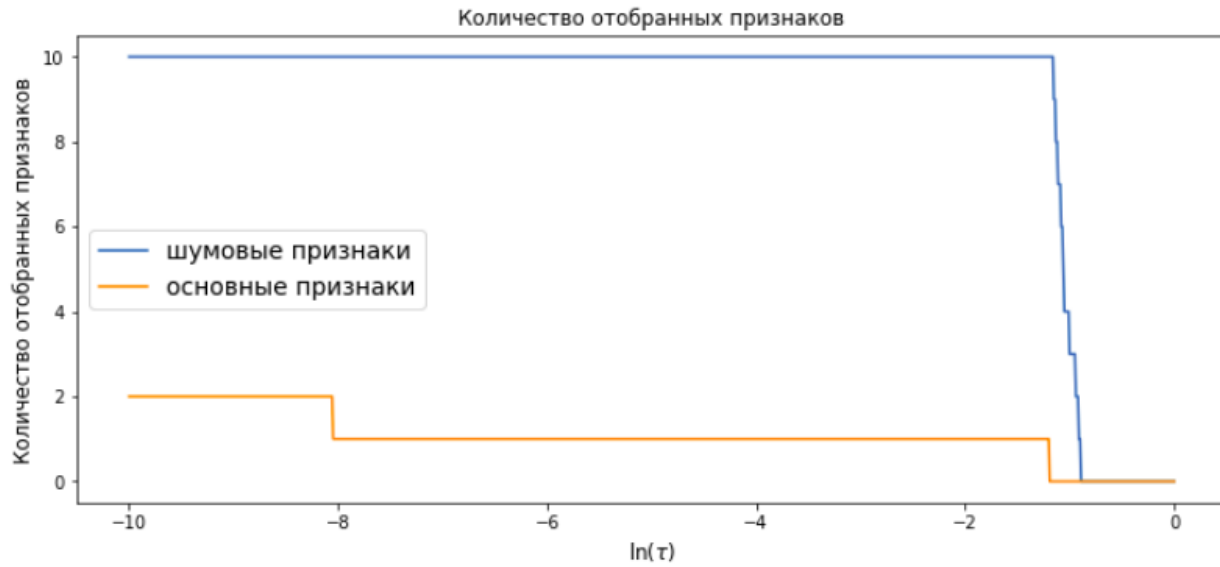


Рис. 1 результаты первого эксперимента

### 4.3 Стабильность модели

Стабильностью модели будем называть отношение  $\lambda_{\max}(\mathbf{X}^T \mathbf{X}) / \lambda_{\min}(\mathbf{X}^T \mathbf{X})$ , где  $\lambda_{\min}$  и  $\lambda_{\max}$  — минимальное и максимальное собственные числа соответствующей матрицы.

В статьях [8, 9], стабильность модели имеет самостоятельную ценность и наряду с качеством прогноза на тестовой выборке определяет решение о превосходстве одного метода отбора признаков над другим. В данной работе предлагается рассматривать стабильность как априорное знание, которое указывает на то, что априори мы считаем, что выборки с меньшим числом обусловленности на множестве активных признаков появляются чаще в рассматриваемой задаче, чем выборки с большим числом обусловленности. При отсутствии такого знания стабильность стоит рассматривать в контексте повышения качества прогноза: если низкая стабильность модели ведет к снижению качества прогноза, стоит добавить штраф за низкую стабильность модели. А если качество не снижается, а повышается при уменьшении стабильности, то не стоит отдавать предпочтение менее качественной, но более стабильной модели.

Обозначим через  $\mathbf{X}(\mathbf{w})$  сужение матрицы признаков на множество признаков  $j : w_j \neq 0$ . Примером соответствующей задачи оптимизации, где есть априорное знание о том, что низкое число обусловленности более предпочтительно, является задача

$$\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \tau \lambda_{\max}(\mathbf{X}(\mathbf{w})^T \mathbf{X}(\mathbf{w})) / \lambda_{\min}(\mathbf{X}(\mathbf{w})^T \mathbf{X}(\mathbf{w})) \rightarrow \min_{\mathbf{w}},$$

что соответствует показательному распределению на числе обусловленности активных признаков выборки с гиперпараметром  $\tau$ . В таком виде задача является слишком трудной для оптимизации и требует перебора наборов активных признаков, например, с помощью генетического алгоритма.

### 4.4 Второй эксперимент

Рассмотрим выборку  $(\mathbf{X}, \mathbf{y})$  в признаковом пространстве размерности  $n$ . Пусть,  $y \sim \mathcal{N}(0, 1)$ ,  $\varepsilon_j \sim \mathcal{N}(0, 1)$  — независимые стандартно нормально распределенные случайные величины (здесь  $j \in \{1, 2, \dots, n\}$ ). Положим,  $x_j = y + \nu \varepsilon_j$ , где  $x_j$  —  $j$ -ый признак, а  $\nu$  — некоторое заранее выбранное значение.

188 Оптимальная модель использует все признаки и усредняет их для получения наилуч-  
 189 шего прогноза целевой переменной:  $\mathbf{w}^* = (1/n, \dots, 1/n)^T$ .

190 Иследуем зависимость числа обусловленности  $\eta$  от параметра  $\nu$ . При каждом зна-  
 191 чении параметра  $\nu$  будем находить число обусловленности соответствующей матрицы и  
 192 усреднять результат по 100 итерациям генерации целевого вектора и выборки. На каждой  
 193 итерации размер целевого вектора и выборки по каждому из признаков полагаем равным  
 194 100.

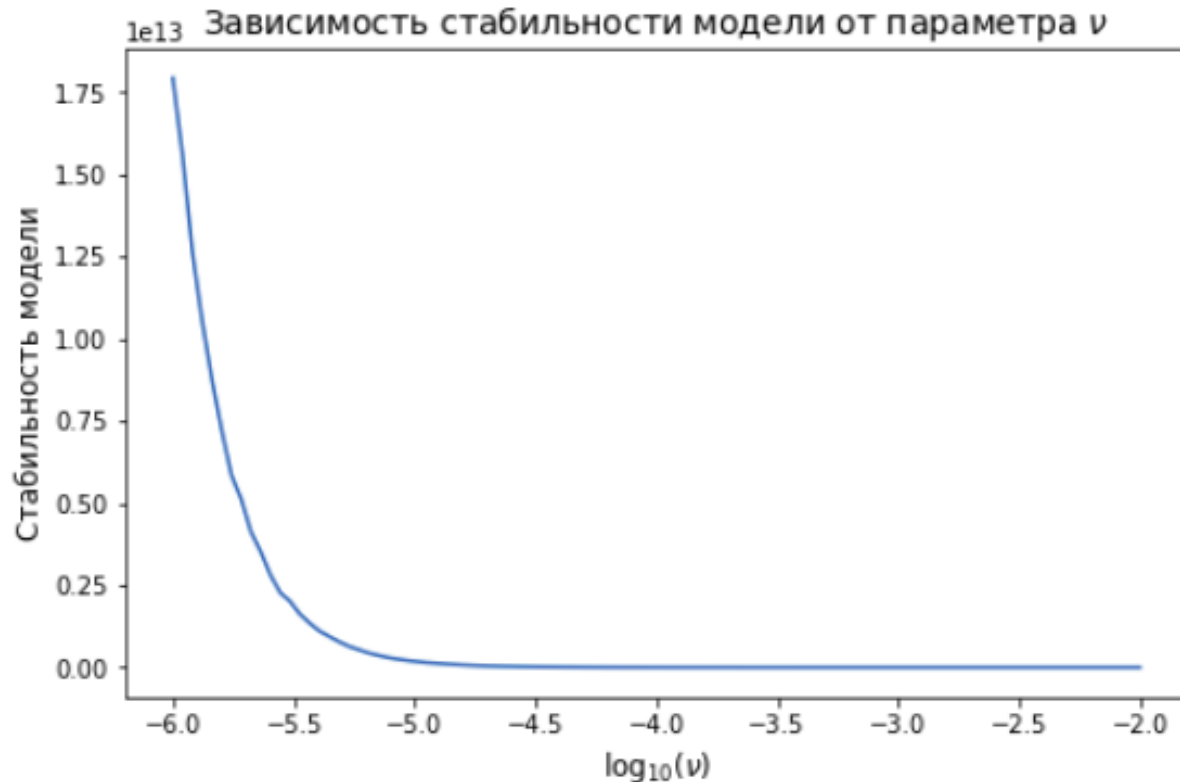


Рис. 2 Зависимость стабильности модели от параметра  $\nu$

195 Получаем, что число обусловленности  $\eta = \lambda_{\max}(\mathbf{X}(\mathbf{w})^T \mathbf{X}(\mathbf{w})) / \lambda_{\min}(\mathbf{X}(\mathbf{w})^T \mathbf{X}(\mathbf{w}))$  при  
 196 малом  $\nu$  может быть большим.

#### 197 4.5 Третий эксперимент

198 Пусть в выборке есть дублирование признаков. Снова положим, что целевой вектор  
 199 имеет стандартное нормальное распределение:  $y \sim \mathcal{N}(0, 1)$ . Признаки  $x_1$  и  $x_2$  получим как  
 200  $x_1 x_2 = y + \nu \varepsilon$ , где  $\varepsilon \sim \mathcal{N}(0, 1)$ .

201 В этом случае число обусловленности равно  $\infty$  и имеется неоднозначность решения,  
 202 минимизирующего  $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$ . Однако все эти решения имеют  $w_1 + w_2 = C = \text{const}$ , а  
 203 потому если на тестовой выборке признаки 1 и 2 останутся идентичными, качество про-  
 204 гноза будет одинаковым независимо от того, какое разбиение этой константы между  $w_1$ ,  
 205  $w_2$ , мы предпочтем. Обычно для того, чтобы сделать решение однозначным, добавляют  
 206 слабую квадратичную регуляризацию на  $\mathbf{w}$  для того, чтобы из всех решений более пред-  
 207 почтительным было то решение, где  $w_1^2 + w_2^2$  минимально, то есть  $w_1 = w_2 = C/2$ .

208 Заметим, что если есть основания полагать, что сильная мультиколлинеарность в обу-  
 209 чающей выборке не будет продолжена в тестовой (см. [2]), то специальная обработка



мультиколлинеарности приобретает важность, а конкретный вид поправок зависит от предположений об эволюции корреляций.

## 5 Заключение

В данной работе был проведен анализ QPFS, который осуществляет отбор признаков на основе метода квадратичного программирования. Была показана связь метода QPFS с Lasso-моделью для линейной регрессии, и теоретически и экспериментально подтверждены недостатки метода QPFS на синтетических данных. Рассмотрены возможности применения числа обусловленности как априорного знания в рассматриваемой модели.

В дальнейшем планируется рассмотреть возможности применения байесовского подхода к методу QPFS для задачи отбора признаков.

## Литература

- [1] Ronald Askin. Multicollinearity in regression: Review and examples. *Journal of Forecasting*, 1(3):281–292, 1982.
- [2] David Belsley. Collinearity and forecasting. *Journal of Forecasting*, 3:183 – 196, 04 1984.
- [3] Verónica Bolón-Canedo, Noelia Sánchez-Marroño, and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34, 03 2012.
- [4] Pablo Estevez, Michel Tesmer, Claudio Perez, and Jacek Zurada. Normalized mutual information feature selection. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 20:189–201, 02 2009.
- [5] Pedram Ghamisi and Jon Benediktsson. Feature selection based on hybridization of genetic algorithm and particle swarm optimization. *Geoscience and Remote Sensing Letters, IEEE*, 12:309–313, 02 2015.
- [6] Isabelle Guyon and André Elisseeff. An introduction of variable and feature selection. *J. Machine Learning Research Special Issue on Variable and Feature Selection*, 3:1157 – 1182, 01 2003.
- [7] Nuhu Ibrahim, H.A. Hamid, Shuzlina Rahman, and Simon Fong. Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science and Technology*, 26:329–340, 01 2018.
- [8] Aleksandr Katrutsa and Vadim Strijov. Stresstest procedure for feature selection algorithms. *Chemometrics and Intelligent Laboratory Systems*, 142, 02 2015.
- [9] Aleksandr Katrutsa and Vadim V. Strijov. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Syst. Appl*, 76:1–11, 2017.
- [10] Edward Leamer. Multicollinearity: A bayesian interpretation. *The Review of Economics and Statistics*, 55(3):371–80, 1973.
- [11] David Madigan and Greg Ridgeway. Discussion of "least angle regression" by efron et al. *Ann Stat*, 32, 07 2004.
- [12] Ranjit Paul. Multicollinearity: Causes, effects and remedies. 04 2021.
- [13] Irene Rodriguez-Lujan, Ramón Huerta, Charles Elkan, and Carlos Cruz. Quadratic programming feature selection. *Journal of Machine Learning Research*, 11:1491–1516, 04 2010.
- [14] Ron Snee. Regression diagnostics: Identifying influential data and sources of collinearity, book review. *Journal of Quality Technology*, 01 1980.
- [15] Vadim Strijov, Katya Krymova, and Gerhard-Wilhelm Weber. Evidence optimization for consequently generated models. *Mathematical and Computer Modelling*, 57, 01 2013.
- [16] Robert Tibshirani. Regression shrinkage selection via the lasso. *Journal of the Royal Statistical Society Series B*, 73:273–282, 06 2011.

- 254 [17] Konstantin Vorontsov. Combinatorial probability and the tightness of generalization bounds.  
255 *Pattern Recognition and Image Analysis*, 18:243–259, 06 2008.

256 *Поступила в редакцию*