

**Цель исследования:** создание процедуры тестирования методов выбора признаков и разработка критерия сравнения методов с точки зрения наличия мультиколлинеарных признаков в множестве отобранных признаков.

**Проблема:** методы выбора признаков могут в качестве решения доставлять подмножество признаков, содержащее мультиколлинеарные признаки.

**Задача:** предложить алгоритм тестирования методов выбора признаков, который

- ранжирует методы выбора признаков;
- определяет количество мультиколлинеарных признаков в множестве отобранных признаков.

# Задача выбора оптимального набора признаков

Задана выборка  $\mathcal{D} = \{(\mathbf{X}, \mathbf{y})\}$ .

$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_n]$  — матрица плана,  $j \in \mathcal{J}$ .

$\mathbf{y} \in \mathbb{R}^m$  — целевой вектор.

Принята модель  $\mathbf{y} = \mathbf{f}(\mathbf{w}, \mathbf{X}) + \boldsymbol{\varepsilon} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$ , где  $\mathbf{w} \in \mathbb{W}$  — пространство допустимых параметров и  $\boldsymbol{\varepsilon}$  — вектор регрессионных остатков.

Задача выбора активного набора индексов:

$$\mathcal{A}^* = \arg \min_{\mathcal{A} \subset \mathcal{J}} S(\mathcal{A} | \mathbf{w}^*, \mathcal{D}_{\mathcal{L}}),$$

где  $\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w} | \mathcal{D}_{\mathcal{L}}, \mathcal{A})$  и  $S = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$ .

Зададим множество методов выбора признаков  $\mathfrak{M} = \{\text{Lasso}, \text{LARS}, \text{Stepwise}, \text{ElasticNet}, \text{Ridge}\}$ .

# Определения

## Определение

Активным набором индексов  $\mathcal{A}_i$  будем называть множество индексов  $\mathcal{A}_i \subseteq \mathcal{J}$ , такое что  $m_i : \mathcal{J} \rightarrow \mathcal{A}_i$ ,  $m_i \in \mathfrak{M}$ .

## Определение

Назовём набор признаков мультиколлинеарным, если найдутся такие коэффициенты  $a_\ell$ ,  $\ell \in \mathcal{B}$  и достаточно малое  $\delta > 0$ , что

$$\left\| \chi_j - \sum_{\ell \in \mathcal{B}} a_\ell \chi_\ell \right\|_2^2 < \delta, \quad j \notin \mathcal{B}.$$

## Определение

Назовём признаки  $\chi_i, \chi_j$  коррелирующими, если найдётся достаточно малое  $\delta_{ij} > 0$  такое, что:

$$\|\chi_i - \chi_j\|_2^2 < \delta_{ij}.$$

# Выборки для тестирования методов выбора признаков

Неадекватная  
коррелирующая:

$$\langle \mathbf{y}, \boldsymbol{\chi}_j \rangle = 0, \quad j \in \mathcal{J};$$

$$\left\| \boldsymbol{\chi}_i - \sum_{l \in \mathcal{B}} \alpha_l \boldsymbol{\chi}_l \right\|_2^2 < \delta,$$

где  $i \in \mathcal{J}$ ,  $i \notin \mathcal{B} \subset \mathcal{J}$ ;

$$\mathcal{J} = \mathcal{P}_y \cap \mathcal{C}_f.$$

Адекватная случайная:

$$\mathcal{J} = \mathcal{R}; \quad \|\mathbf{y} - \boldsymbol{\chi}_i\|_2^2 < \delta;$$

$$\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_r \sim U[0, 1]^r.$$

Адекватная коррелирующая:

$$\langle \boldsymbol{\chi}_i, \boldsymbol{\chi}_j \rangle = 0, \quad i, j \in \mathcal{P}_f;$$

$$\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|_2^2 < \delta_{ij}, \quad i \in \mathcal{P}_f, j \in \mathcal{C}_f;$$

$$\mathbf{y} = \sum_{j \in \mathcal{P}_f} a_j \boldsymbol{\chi}_j;$$

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

Адекватная избыточная:

$$\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|_2^2 < \delta_{ij}, \quad i, j \in \mathcal{J};$$

$$\|\mathbf{y} - \boldsymbol{\chi}_j\|_2^2 < \delta, \quad j \in \mathcal{J};$$

$$\mathcal{J} = \mathcal{C}_y.$$

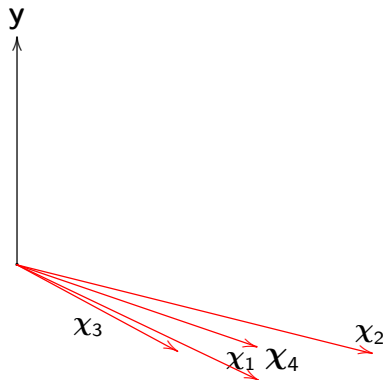
# Неадекватная коррелирующая выборка

$$\langle \mathbf{y}, \boldsymbol{\chi}_j \rangle = 0, \quad j \in \mathcal{J};$$

$$\left\| \boldsymbol{\chi}_i - \sum_{l \in \mathcal{B}} \alpha_l \boldsymbol{\chi}_l \right\|_2^2 < \delta,$$

где  $i \in \mathcal{J}$ ,  $i \notin \mathcal{B} \subset \mathcal{J}$ ;

$$\mathcal{J} = \mathcal{P}_y \cap \mathcal{C}_f.$$

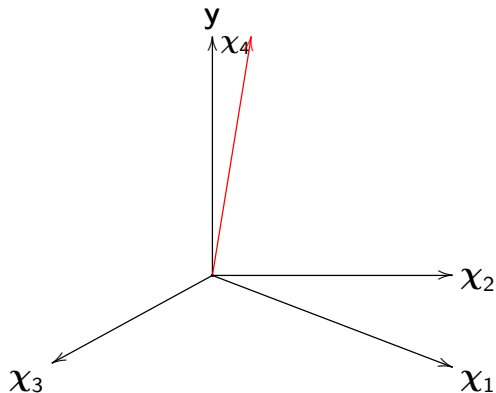


# Адекватная случайная выборка

$$\mathcal{J} = \mathcal{R};$$

$$\chi_1, \dots, \chi_r \sim U[0, 1]^r;$$

$$\|y - \chi_i\|_2^2 < \delta.$$

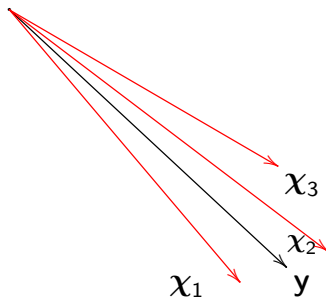


# Адекватная избыточная выборка

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 < \delta_{ij}, \quad i, j \in \mathcal{J};$$

$$\|\boldsymbol{y} - \boldsymbol{x}_j\|_2^2 < \delta, \quad j \in \mathcal{J};$$

$$\mathcal{J} = \mathcal{C}_y.$$



# Адекватная коррелирующая выборка

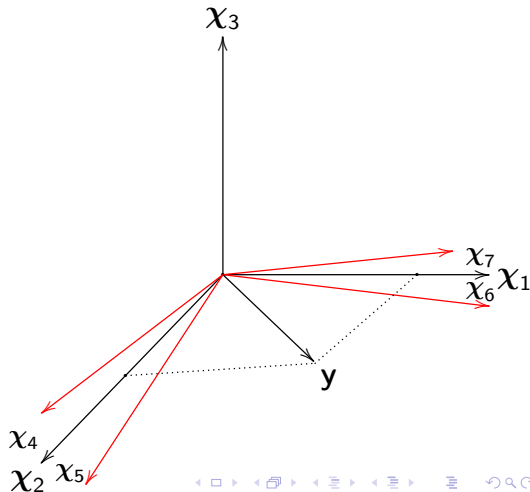
$$\langle \chi_i, \chi_j \rangle = 0, \quad i, j \in \mathcal{P}_f;$$

$$\|\chi_i - \chi_j\|_2^2 < \delta_{ij},$$

$$i \in \mathcal{P}_f, j \in \mathcal{C}_f;$$

$$\mathbf{y} = \sum_{j \in \mathcal{P}_f} a_j \chi_{j_i}$$

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





# Структура тестовых выборок

Структуру тестовой выборки задают следующие множества:

- 1) множество ортогональных признаков  $\chi_j$  с индексами  $j$  из множества  $\mathcal{P}_f$ ;
- 2) множество признаков  $\chi_j$  ортогональных целевому вектору  $y$  с индексами  $j$  из множества  $\mathcal{P}_y$ ;
- 3) множество мультиколлинеарных признаков  $\chi_j$  с индексами  $j$  из множества  $\mathcal{C}_f$ ;
- 4) множество признаков  $\chi_j$ , коррелирующих с целевым вектором, с индексами  $j$  из множества  $\mathcal{C}_y$ ;
- 5) множество случайных признаков  $\chi_j$  с индексами из множества  $\mathcal{R}$ .

Параметр мультиколлинеарности  $k$ : при  $k = 1$  признаки коллинеарны, при  $k = 0$  — ортогональны.

# Критерий сравнения методов выбора признаков

Пусть  $s_0$  — предельно допустимое значение функции ошибки  $S(\mathcal{J}|\mathbf{w}, \mathcal{D})$ , максимальная мощность  $h$  множества индексов признаков  $\mathcal{J}_h \subseteq \mathcal{A}$ , при удалении которого значение функции ошибки  $S$  не превосходит  $s_0$ :

$$h = \arg \max_{S(\mathcal{J}_h|\mathbf{w}_h, \mathcal{D}) \leq s_0} |\mathcal{J}_h|.$$

Получим  $d$  — количество избыточных признаков, удаление которых приводит к ошибке, не превышающей  $s_0$ :

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

## Критерий

*Чем больше значение  $d$ , тем более избыточно множество признаков  $\mathcal{A}_i$  — решение получаемое методом выбора признаков  $m_i$  и тем хуже метод выбора признаков  $m_i$ .*

## Цели эксперимента:

- показать на различных выборках отсутствие универсального, оптимального в смысле введённого критерия, метода выбора признаков;
- показать зависимость количества избыточных признаков  $d$  от критической ошибки  $s_0$  для рассматриваемых методов выбора признаков;
- показать зависимость VIF от параметра мультиколлинеарности  $k$  для множества отобранных признаков.

Параметры экспериментов:

$m = 1000$ ,  $n = 50$ ,  $k = 0.2$  или  $k = 0.8$ ,  $s_0 = 0.5$ .

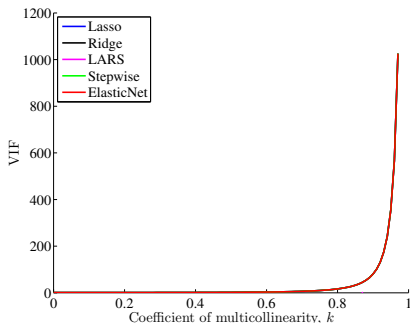
# Зависимость VIF от $k$ для неадекватных коррелирующих выборок

Ни один из рассматриваемых методов выбора признаков не решает проблему мультиколлинеарности.

## Определение

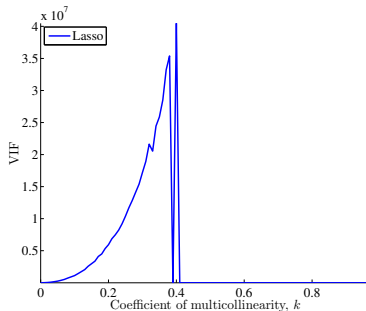
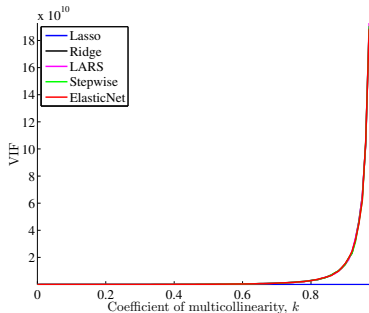
$VIF_j = \frac{1}{1-R_j^2}$ , где  $R_j^2$  — коэффициент детерминации, где целевой вектор —  $j$ -ый признак,  $j \in \mathcal{A}$ ,  $\mathcal{J} = \mathcal{A} \setminus \{j\}$ .

$$VIF = \max_{j \in \mathcal{A}} VIF_j$$

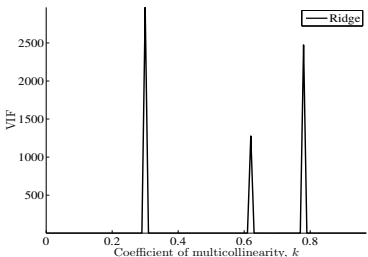
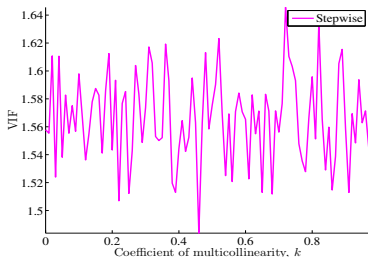
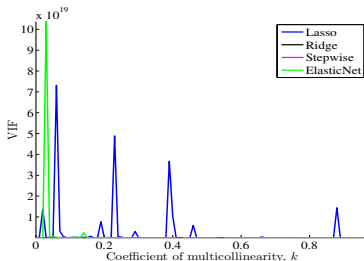
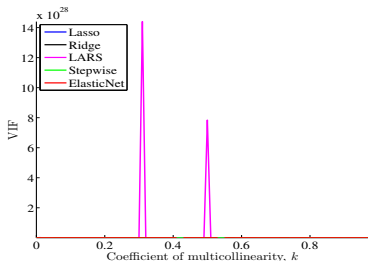


# Зависимость VIF от $k$ для адекватных избыточных выборок

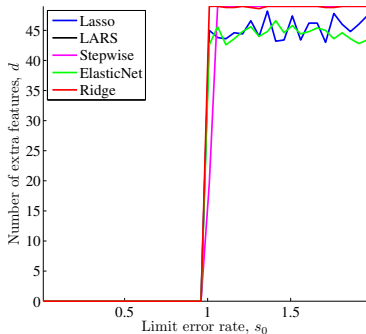
Проблему мультиколлинеарности решает метод Lasso.



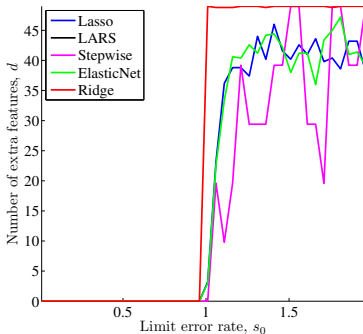
Проблему мультиколлинеарности решает метод Stepwise.



# Зависимость $d$ от $s_0$ для неадекватных выборок

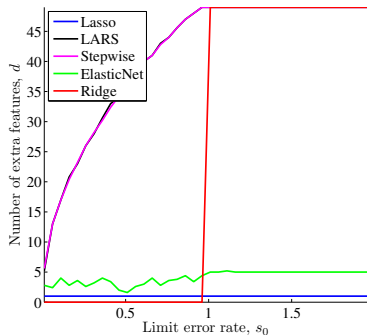


(a)  $k = 0.2$

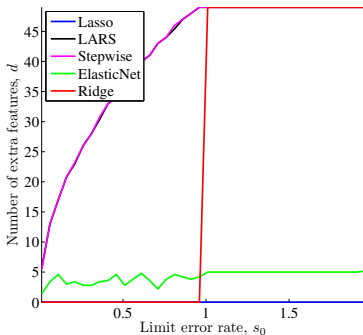


(b)  $k = 0.8$

# Зависимость $d$ от $s_0$ для адекватных избыточных выборок



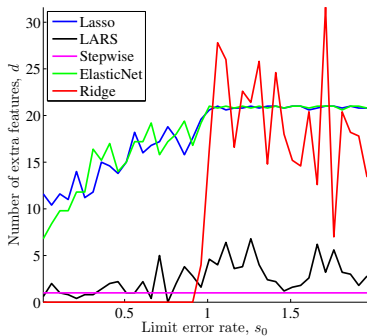
(a)  $k = 0.2$



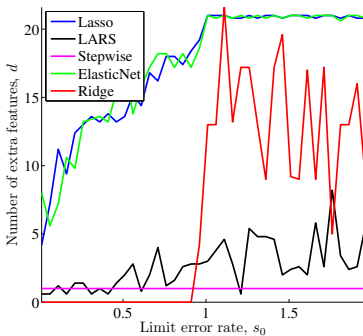
(b)  $k = 0.8$



# Зависимость $d$ от $s_0$ для адекватных коррелирующих выборок

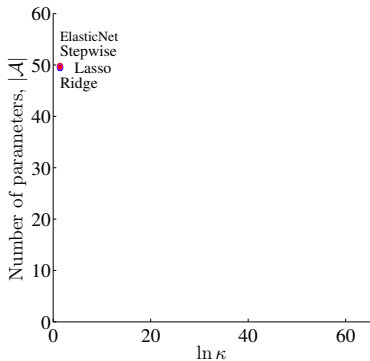


(a)  $k = 0.2$

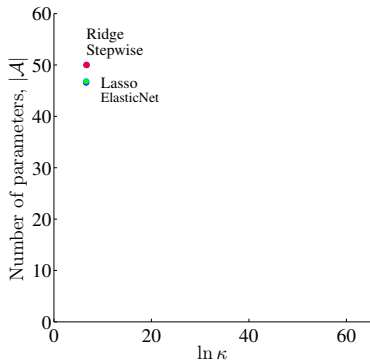


(b)  $k = 0.8$

# Сложность и устойчивость моделей для неадекватных коррелирующих выборок

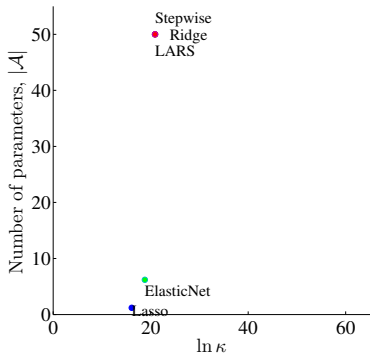


(a)  $k = 0.2$

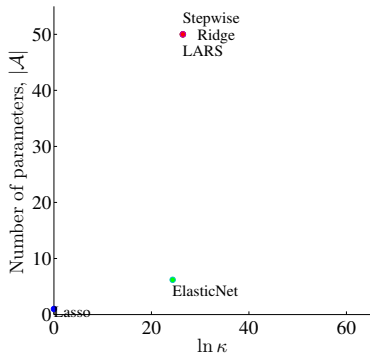


(b)  $k = 0.8$

# Сложность и устойчивость моделей для адекватных избыточных выборок

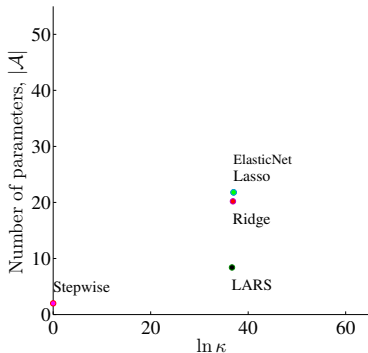


(a)  $k = 0.2$

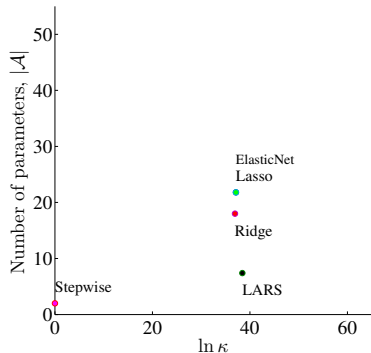


(b)  $k = 0.8$

# Сложность и устойчивость моделей для адекватных коррелирующих выборок



(a)  $k = 0.2$



(b)  $k = 0.8$