

# Цели исследования

## Цель работы

Предложить метод отбора признаков, учитывающий взаимное расположение признаков и целевого вектора.

## Проблема

Методы отбора признаков дают избыточное подмножество мультикоррелирующих признаков.

## Метод решения

Использование постановки задачи квадратичного программирования для получение оптимального подмножества признаков.

# Постановка задачи

Пусть  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$  — матрица плана,  
 $\mathbf{y} \in \mathbb{R}^m$  — целевой вектор.

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathbb{R}^n} S(\mathbf{w}, \mathcal{A} | \mathbf{X}, \mathbf{y}, \mathbf{f}),$$

где  $S$  — функция ошибки,  $\mathbf{f}$  — модель,  $\mathcal{A}$  — множество активных признаков.

$S(\mathbf{w}, \mathcal{A} | \mathbf{X}, \mathbf{y}, \mathbf{f}) = \|\mathbf{f}(\mathbf{X}, \mathcal{A}, \mathbf{w}) - \mathbf{y}\|_2^2$  и  $\mathbf{f}(\mathbf{X}, \mathcal{A}, \mathbf{w}) = \mathbf{X}_{\mathcal{A}} \mathbf{w}$ ,  
 $\mathbf{X}_{\mathcal{A}} \in \mathbb{R}^{m \times |\mathcal{A}|}$ .

$$\mathcal{A}^* = \arg \min_{\mathcal{A} \subseteq \mathcal{J}} Q(\mathcal{A} | \mathbf{X}, \mathbf{y}),$$

где  $Q : \mathcal{A} \rightarrow \mathbb{R}$  критерий качества подмножества индексов признаков  $\mathcal{A} \subseteq \mathcal{J} = \{1, \dots, n\}$ .

$$\mathbf{a}^* = \arg \min_{\mathbf{a} \in \mathbb{B}^n} Q(\mathbf{a} | \mathbf{X}, \mathbf{y}),$$

где  $Q : \mathbb{B}^n \rightarrow \mathbb{R}$ .

Типы признаков:

- информативные — существенно влияют на точность приближения целевого вектора
- шумовые — не влияют на точность приближения целевого вектора
- мультиколлинеарные — существует линейная зависимость между признаками, снижают устойчивость модели

# Задача квадратичного программирования

$$Q(\mathbf{a}|\mathbf{Q}, \mathbf{b}) = \mathbf{a}^T \mathbf{Q} \mathbf{a} - \mathbf{b}^T \mathbf{a},$$

где матрица  $\mathbf{Q} \in \mathbb{R}^{n \times n}$  и вектор  $\mathbf{b} \in \mathbb{R}^n$ .

- $\mathbf{Q} = [q_{ij}] = \text{Sim}(\chi_i, \chi_j)$  — похожесть между признаками  $i$  и  $j$
- $\mathbf{b} = [b_i] = \text{Rel}(\chi_i)$  — релевантность признака  $i$  целевому вектору

Задача квадратичного программирования для отбора признаков:

$$\mathbf{a}^* = \arg \min_{\mathbf{a} \in \mathbb{B}^n} \mathbf{a}^T \mathbf{Q} \mathbf{a} - \mathbf{b}^T \mathbf{a}.$$

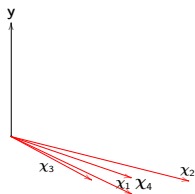
- Корреляция

$$q_{ij} = \left| \frac{\text{cov}(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{\text{Var}(\mathbf{x}_i)\text{Var}(\mathbf{x}_j)}} \right| \quad b_i = \left| \frac{\text{cov}(\mathbf{x}_i, \mathbf{y})}{\sqrt{\text{Var}(\mathbf{x}_i)\text{Var}(\mathbf{y})}} \right|$$

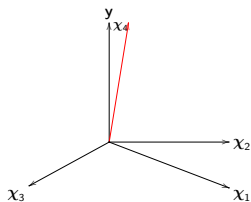
- Взаимная информация

$$I(\mathbf{x}_i, \mathbf{x}_j) = \int \int p(\mathbf{x}_i, \mathbf{x}_j) \log \frac{p(\mathbf{x}_i, \mathbf{x}_j)}{p(\mathbf{x}_i)p(\mathbf{x}_j)} d\mathbf{x}_i d\mathbf{x}_j.$$

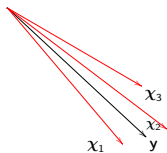
# Тестовые выборки



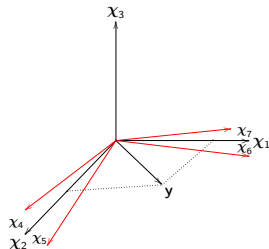
Неадекватная коррелирующая



Адекватная случайная



Адекватная избыточная



Адекватная коррелирующая

# Вычислительный эксперимент

## Цель эксперимента

Проверить работоспособность предложенного метода и сравнить с существующими методами отбора признаков

Название	Формула	Значение
VIF	$\text{VIF} = \max_{j \in \mathcal{A}} \frac{1}{1 - R_j^2}$	Показатель мультиколлинеарности
Устойчивость	$R = \ln \frac{\lambda_{\min}}{\lambda_{\max}}$	Показатель устойчивости модели
Точность	$r = \ \mathbf{y} - \mathbf{X}\mathbf{w}\ _2^2$	Норма вектора остатков
Mallow's $C_p$	$C_p = \frac{r_p}{r} - m + 2p$	Баланс между точностью и сложностью
BIC	$\text{BIC} = r + p \log m$	Баланс между точностью и сложностью

Table 6: Evaluation criteria for the diesel NIR spectra dataset

Method	$C_p$	RSS	$R$	VIF	BIC
QP ( $\rho, \rho$ ) ( $\tau = 10^{-9}$ )	-110	$1.37 \cdot 10^{-18}$	-25.7	$6.43 \cdot 10^6$	548.38
Genetic	-110.88	$7.68 \cdot 10^{-30}$	-24	$8.13 \cdot 10^5$	534.19
LARS	$3.22 \cdot 10^{21}$	$2.07 \cdot 10^{-7}$	-28.3	$7.94 \cdot 10^7$	529.47
Lasso	$2.5 \cdot 10^{28}$	1.61	-27.72	$1.03 \cdot 10^{21}$	1712.92
ElasticNet	$2.51 \cdot 10^{28}$	1.61	-27.72	$1.03 \cdot 10^{21}$	1712.92
Stepwise	$3.66 \cdot 10^{29}$	23.56	-36.78	$1.94 \cdot 10^{22}$	1919.23
Ridge	$1.59 \cdot 10^{28}$	1.02	-36.22	$1.07 \cdot 10^{22}$	$1.79 \cdot 10^3$



# Сравнение с другими методами

Параметры:  $m = 1000$ ,  $n = 50$

## Неадекватная коррелирующая выборка

Method	$C_p$	$r$	$R$	VIF	BIC
QP( $\rho$ )	-997	—	—	—	—
LARS	-997	—	—	—	—
Genetic	-997	—	—	—	—
Lasso	-997	1	-6.57	16.6	310.48
Ridge	-997	1	-6.69	16.6	346.39
Stepwise	-997	1.68	-6.69	16.6	347.01
Elastic Net	-997	1	-6.58	16.6	310.48

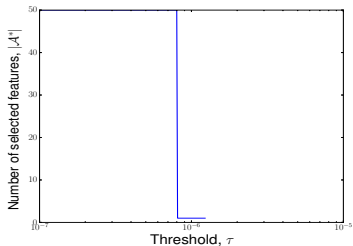
## Адекватная случайная выборка

Method	$C_p$	$r$	$R$	VIF	BIC
QP( $\rho$ )	-997	$1.2 \cdot 10^{-9}$	0	0.24	6.9
Lasso	$7 \cdot 10^6$	$8.50 \cdot 10^{-4}$	0	0.25	6.9
Elastic Net	$8.76 \cdot 10^{-4}$	$8.76 \cdot 10^{-4}$	0	0.25	6.9
Ridge	$7.97 \cdot 10^9$	0.97	0	0.25	7.88
LARS	-997	$1.3 \cdot 10^{-10}$	-0.78	0.32	8.29
Genetic	-997	$1.36 \cdot 10^{-10}$	-3.31	0.9	52.5
Stepwise	-997	$1.33 \cdot 10^{-10}$	-3.36	0.89	53.88

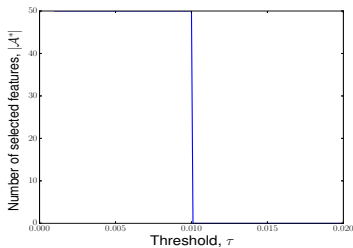
## Адекватная избыточная выборка

Method	$C_p$	$r$	$R$	VIF	BIC
QP( $\rho$ )	-997	$8.5 \cdot 10^{-11}$	0	0.25	6.9
Lasso	$5.16 \cdot 10^8$	$8.5 \cdot 10^{-4}$	0	0.24	6.9
Ridge	$5.9 \cdot 10^{11}$	0.97	-27.13	$2.9 \cdot 10^9$	346.36
Elastic Net	$5.16 \cdot 10^8$	$8.5 \cdot 10^{-4}$	-25.01	$2.5 \cdot 10^9$	41.45
Genetic	-997	$1.67 \cdot 10^{-12}$	-27.11	$2.87 \cdot 10^9$	345.39
Stepwise	-997	$1.73 \cdot 10^{-12}$	-27.13	$2.9 \cdot 10^9$	345.39
LARS	-997	$1.65 \cdot 10^{-12}$	-27.13	$2.9 \cdot 10^9$	345.39

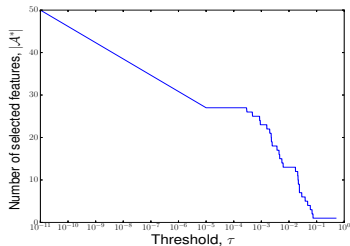
# Результаты на тестовых выборках



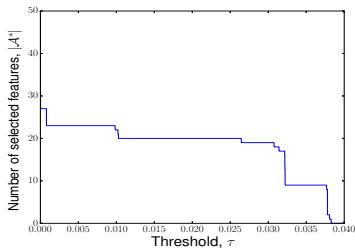
Неадекватная коррелирующая



Адекватная избыточная

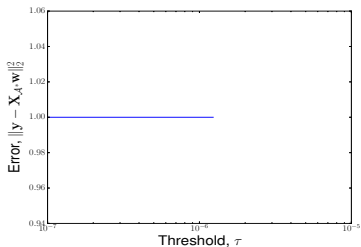


Адекватная случайная

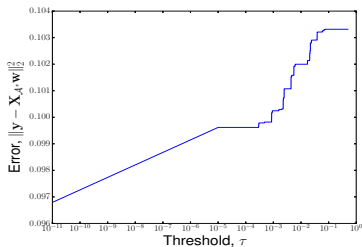


Адекватная коррелирующая

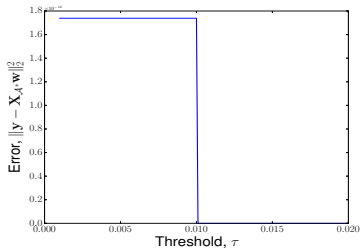
# Результаты на тестовых выборках



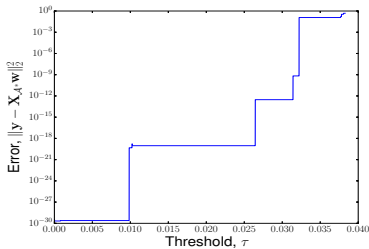
Неадекватная коррелирующая



Адекватная случайная

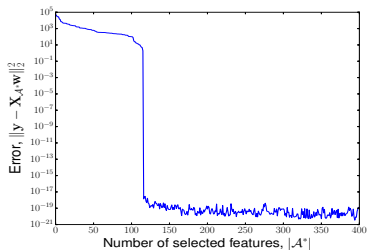


Адекватная избыточная

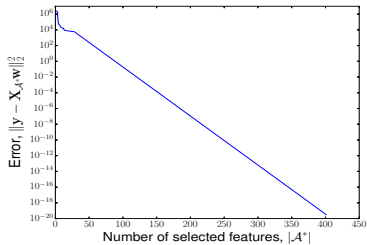


Адекватная коррелирующая

# Результаты на реальных данных



Корреляция



Взаимная информация