Анализ мультиколлинеарности при выборе признаков

А. М. Катруца

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель к.ф.-м.н., н.с. ВЦ РАН В. В. Стрижов

Москва,
2014 г.

Цель исследования

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 $\mathfrak{D} = \{(\mathbf{X}, \mathbf{y})\}$.
$\mathbf{X} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_j, \ldots, \boldsymbol{\chi}_n]$ — 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:

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

где $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w}|\mathfrak{D}_{\mathcal{L}}, \mathcal{A})$ и $S = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$.
Зададим множество методов выбора признаков
$\mathfrak{M} = \{\mathrm{Lasso}, \mathrm{LARS}, \mathrm{Stepwise}, \mathrm{ElasticNet}, \mathrm{Ridge}\}$.

Определения

Определение

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

Определение

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

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

Определение

Назовём признаки $\boldsymbol{\chi}_i, \boldsymbol{\chi}_j$ коррелирующими, если найдётся достаточно малое $\delta_{ij} > 0$ такое, что:

$$\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|_2^2 < \delta_{ij}.$$

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

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

Адекватная случайная:

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

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

Адекватная коррелирующая:

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

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

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

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

Адекватная избыточная:

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

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

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

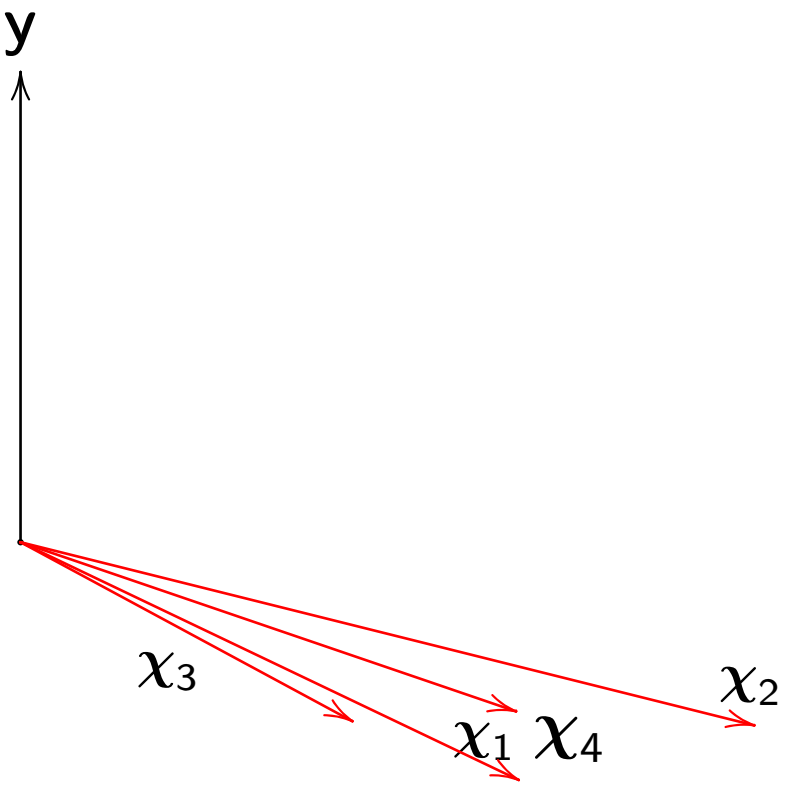
Неадекватная коррелирующая выборка

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

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

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

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

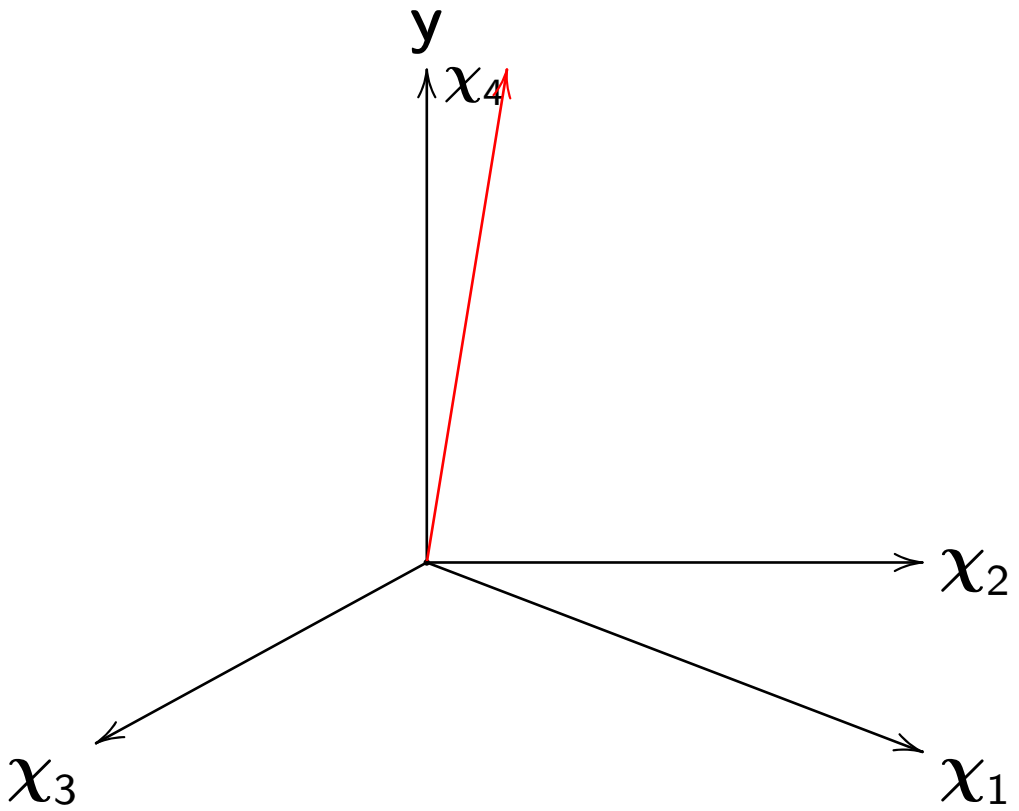


Адекватная случайная выборка

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

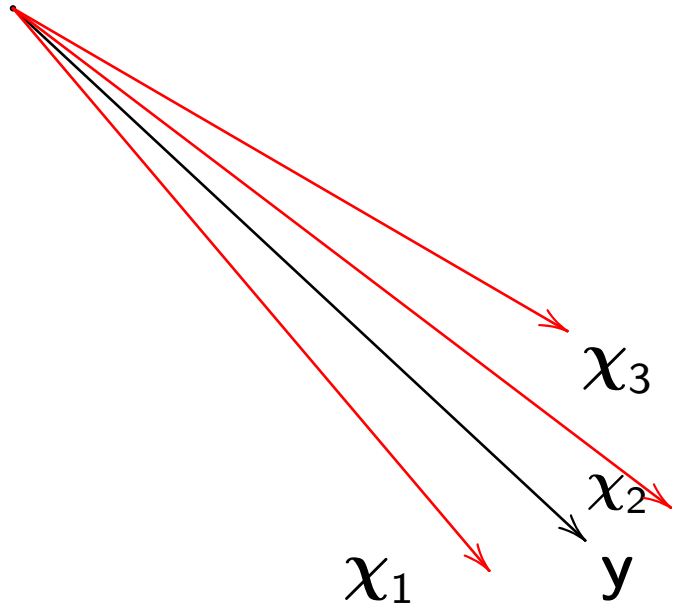
$\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r \sim \mathrm{U}[0,1]^r$;

$\|\mathbf{y} - \boldsymbol{\chi}_i\|_2^2 < \delta$.



Адекватная избыточная выборка

$$\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|_2^2 < \delta_{ij},\ i,j \in \mathcal{J};$$
$$\|\mathbf{y} - \boldsymbol{\chi}_j\|_2^2 < \delta, \quad j \in \mathcal{J};$$
$$\mathcal{J} = \mathcal{C}_y.$$



Адекватная коррелирующая выборка

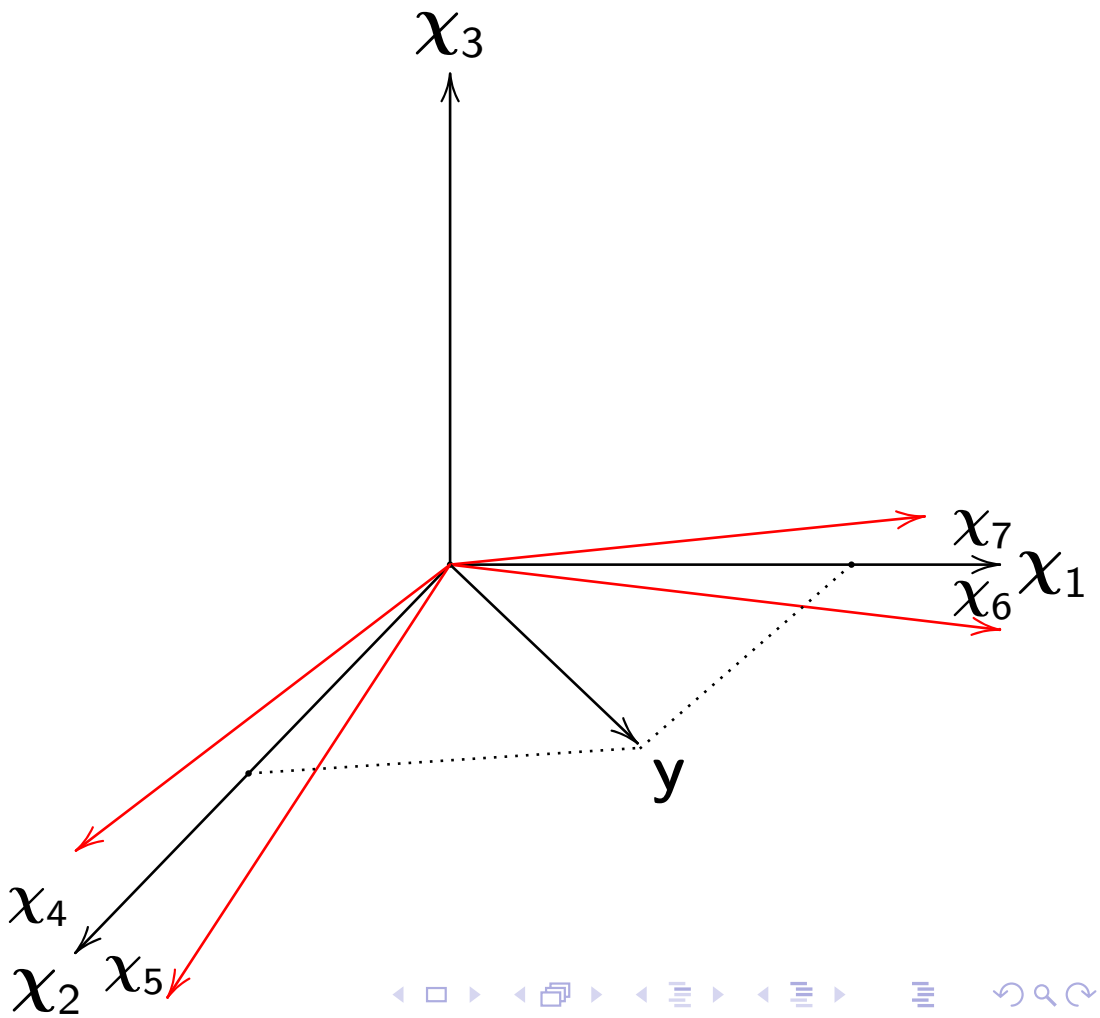
$$\langle \boldsymbol{\chi}_i, \boldsymbol{\chi}_j \rangle = 0, \quad i, j \in \mathcal{P}_f;$$

$$\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|_2^2 < \delta_{ij},$$

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

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

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



Структура тестовых выборок

Структуру тестовой выборки задают следующие множества:

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 $\mathcal{R}$.

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

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

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 $\mathfrak{m}_i$ и тем хуже метод выбора признаков $\mathfrak{m}_i$.

Метод удаления признаков

Индекс обусловленности $\eta_i = \frac{\lambda_{\max}}{\lambda_i}$ в разложении $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}}$, где $\mathbf{U}$ и $\mathbf{V}$ — ортогональные матрицы, а $\mathbf{\Lambda}$ — диагональная с λ_i на диагонали.

Нужно найти такой индекс i^*, что $i^* = \arg\max\limits_{i \in \mathcal{B}_{s-1}} \eta_i$.

Дисперсионная доля $q_{ij} = \frac{v_{ij}^2/\lambda_j^2}{\sum\limits_{j=1}^{n} v_{ij}^2/\lambda_j^2}$, где $[v_{ij}] = \mathbf{V}$.

По найденному максимальному индексу обусловленности i^* находим признак j^*:

$$j^* = \arg\max_{j \in \mathcal{B}_{s-1}} q_{i^* j},$$

который подлежит удалению:

$$\mathcal{B}_s = \mathcal{B}_{s-1} \setminus \{j^*\}.$$

Вычислительный эксперимент

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 $k = 0.8$, $s_0 = 0.5$.

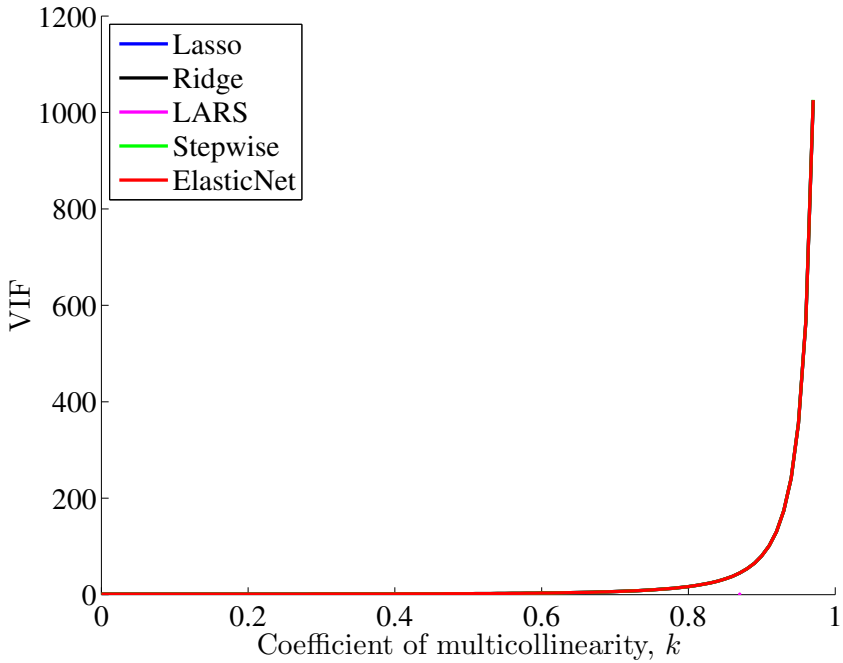
Зависимость VIF от k для неадекватных коррелирующих выборок

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

Определение

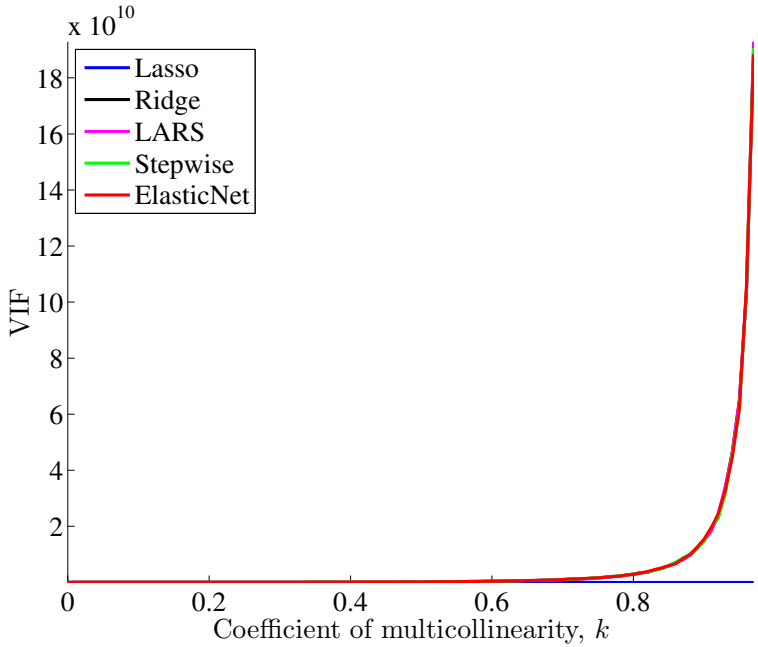
$\mathrm{VIF}_j = \frac{1}{1-R_j^2}$, *где* R_j^2 — *коэффициент детерминации, где целевой вектор — j-ый признак,* $j \in \mathcal{A}, \mathcal{J} = \mathcal{A} \setminus \{j\}$.

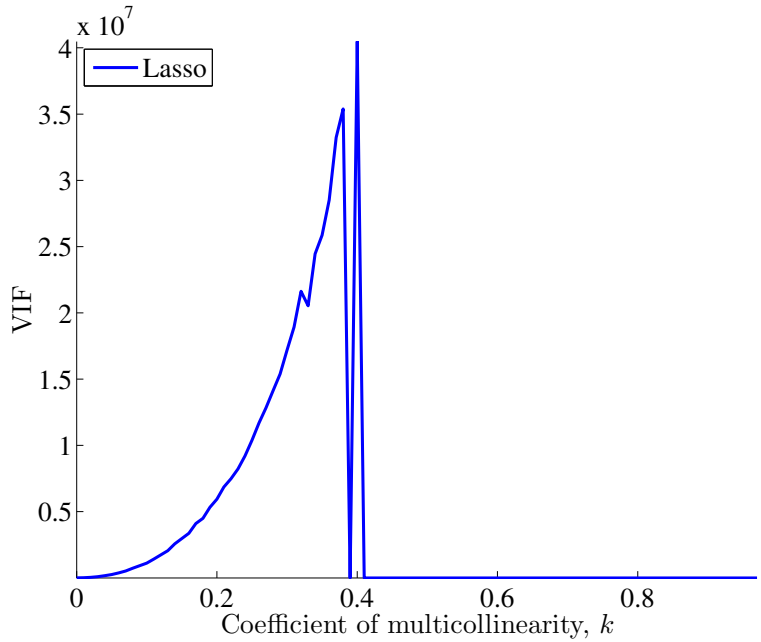
$$\mathrm{VIF} = \max_{j \in \mathcal{A}} \mathrm{VIF}_j$$



Зависимость VIF от k для адекватных избыточных выборок

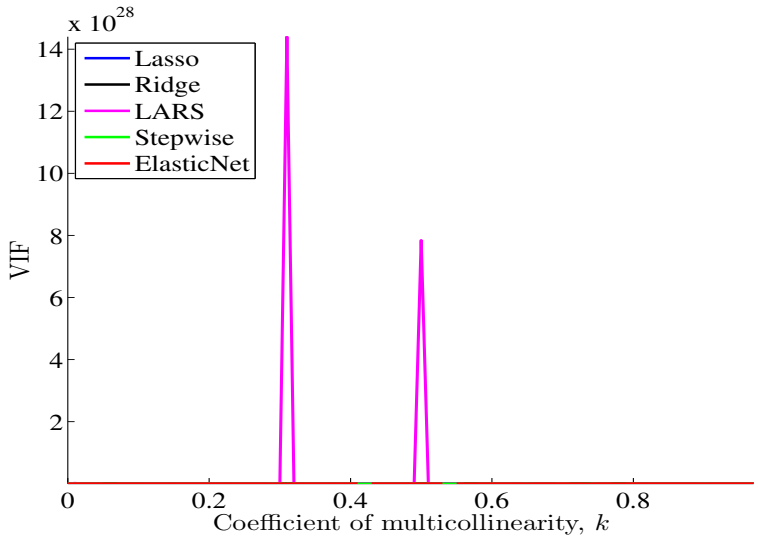
Проблему мультиколлинеарности решает метод Lasso.

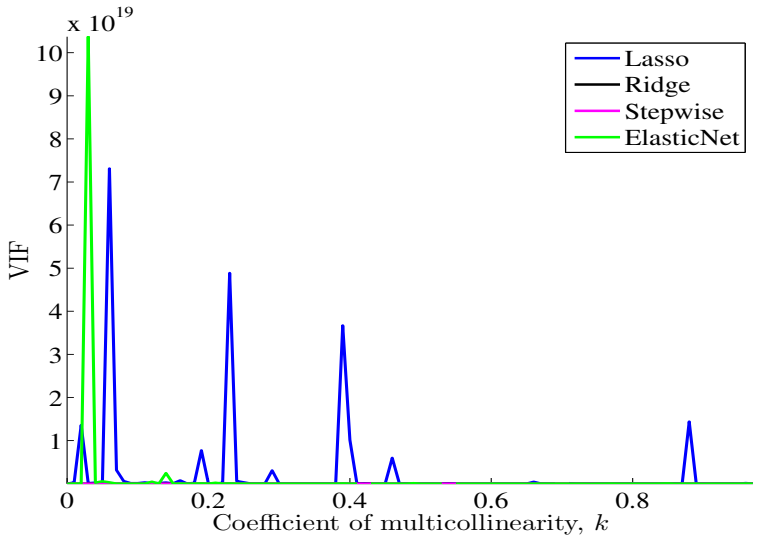


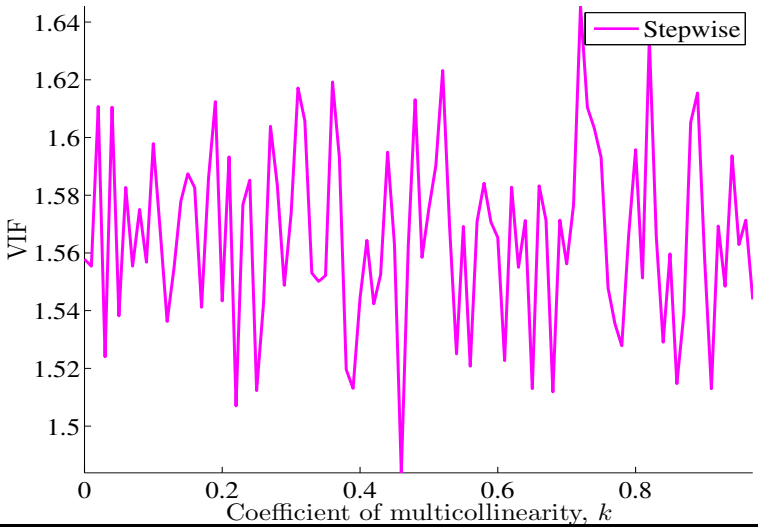


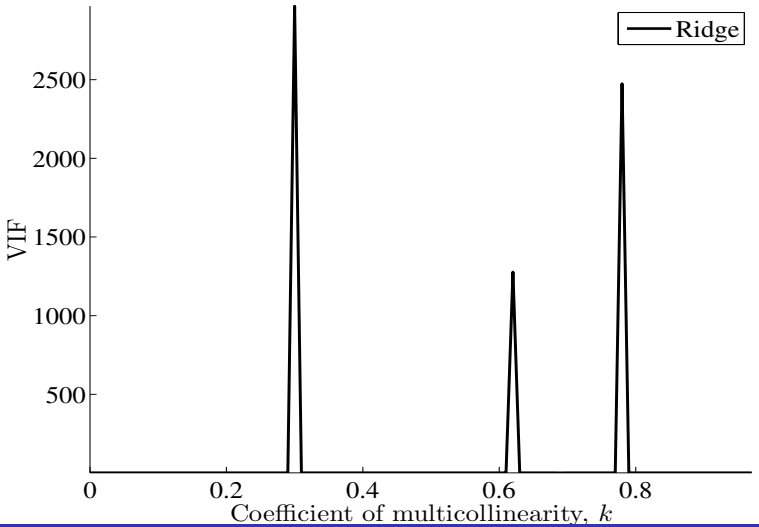
Зависимость VIF от k для адекватных коррелирующих выборок

Проблему мультиколлинеарности решает метод Stepwise.

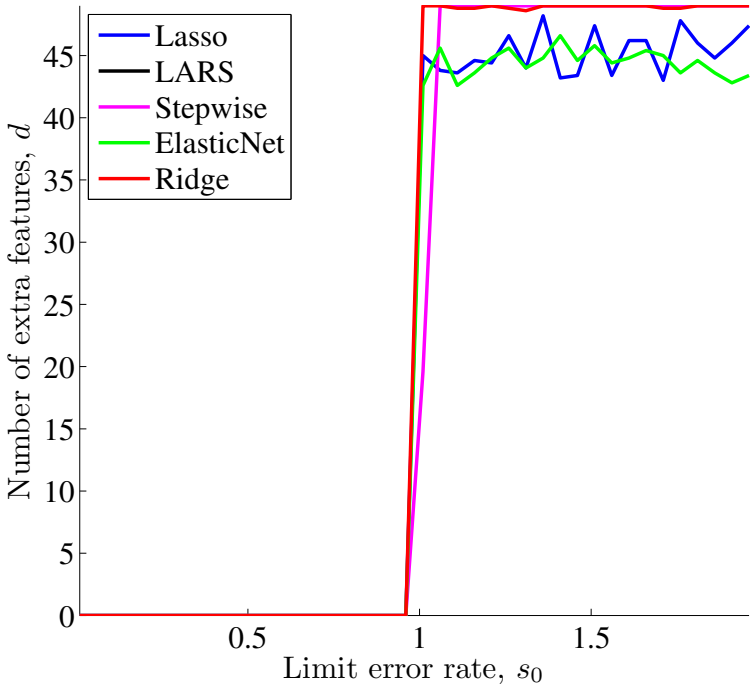




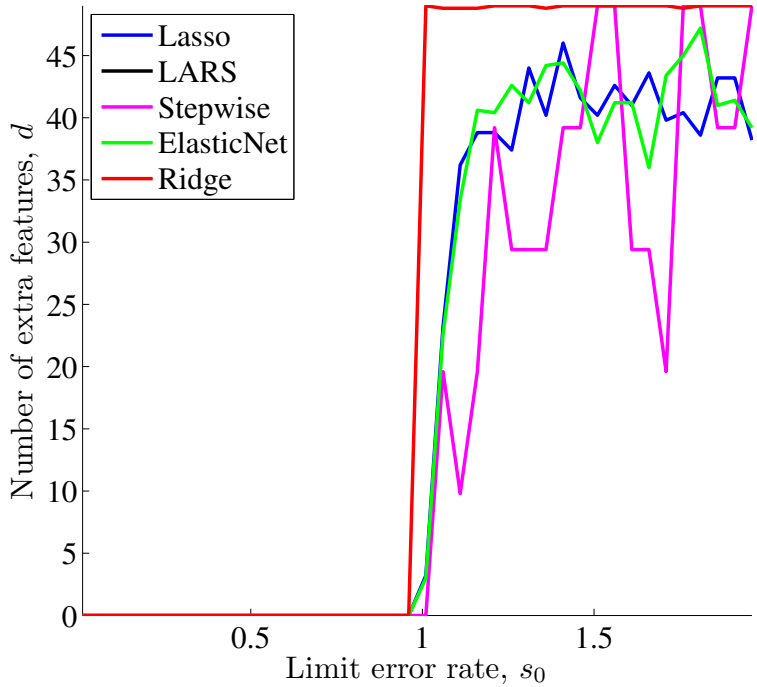




Зависимость d от s_0 для неадекватных коррелирующих выборок

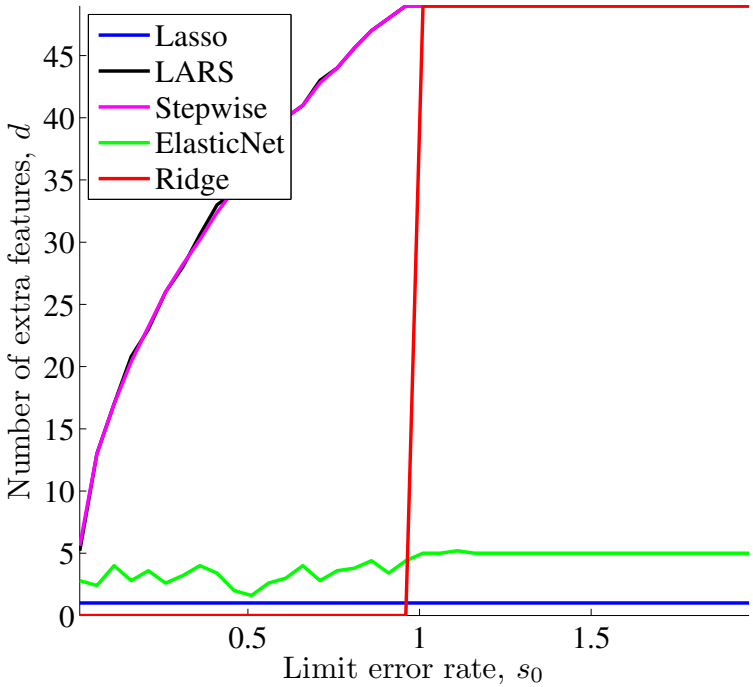


(a) $k = 0.2$

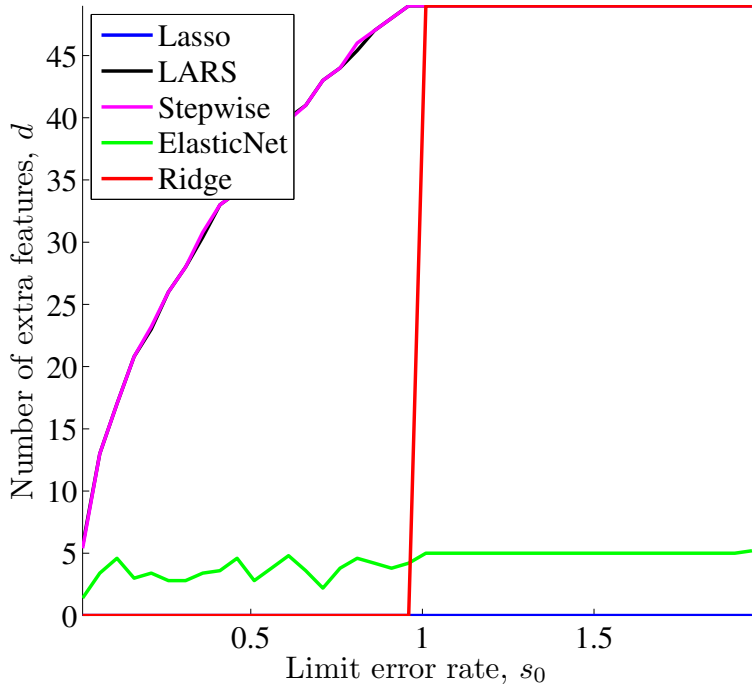


(b) $k = 0.8$

Зависимость d от s_0 для адекватных избыточных выборок

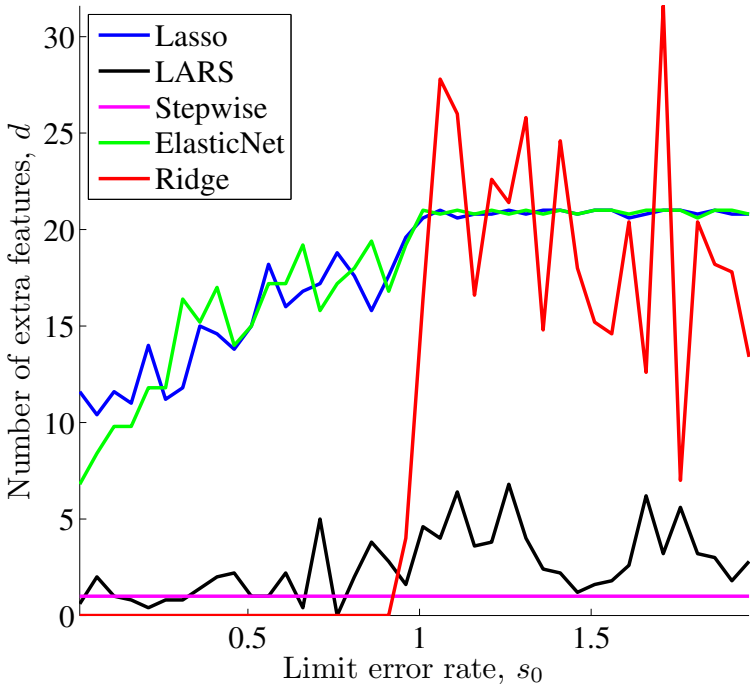


(a) $k = 0.2$

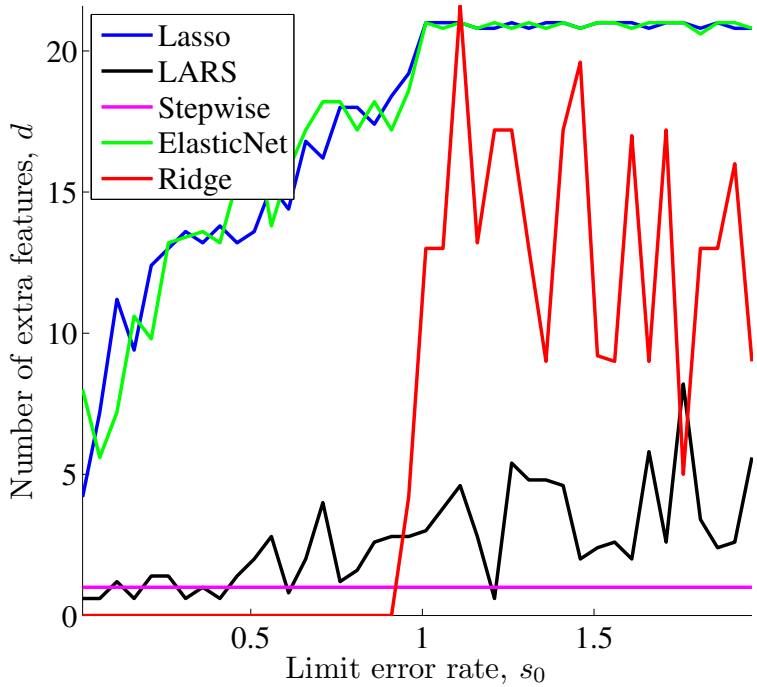


(b) $k = 0.8$

Зависимость d от s_0 для адекватных коррелирующих выборок

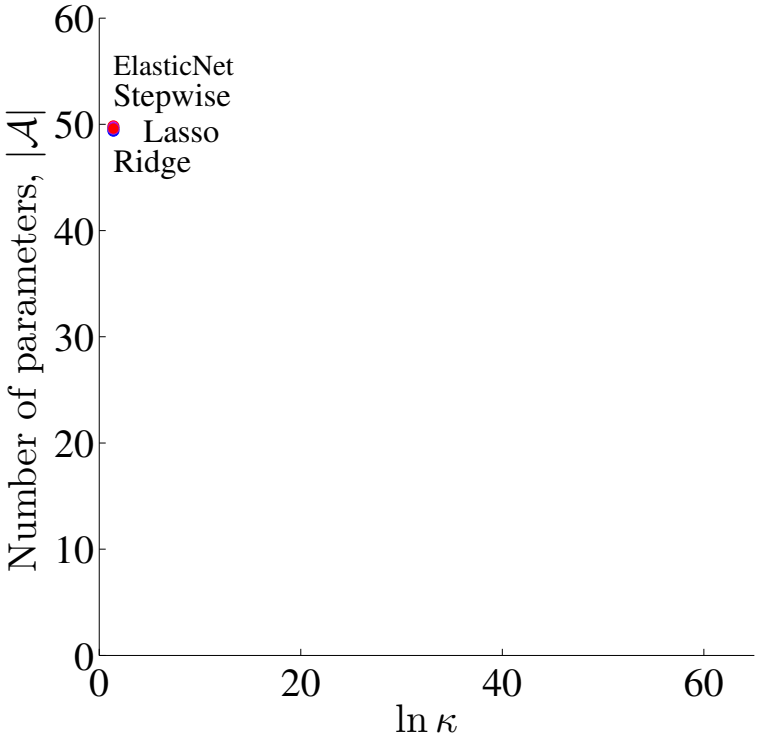


(a) $k = 0.2$

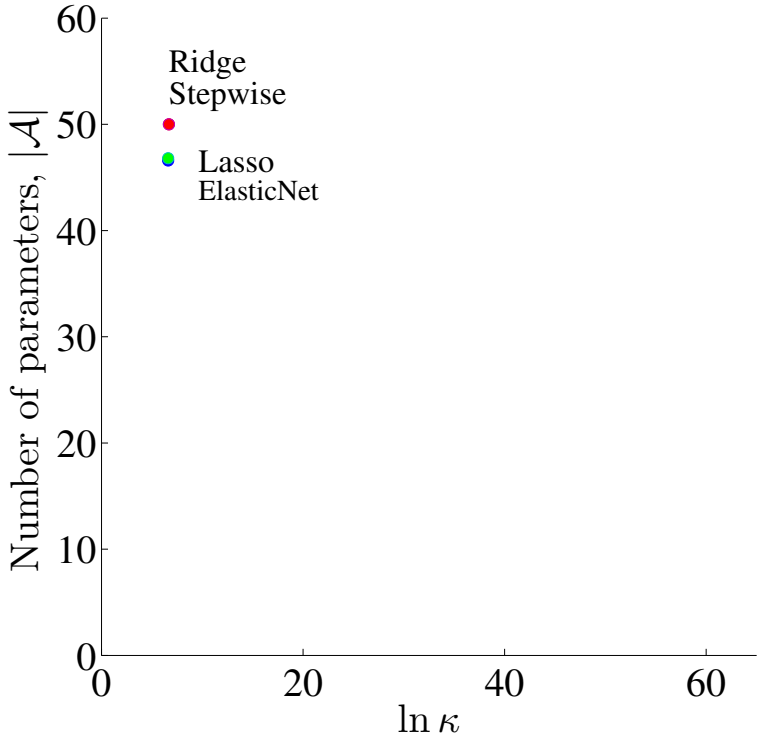


(b) $k = 0.8$

Сложность и устойчивость моделей для неадекватных коррелирующих выборок

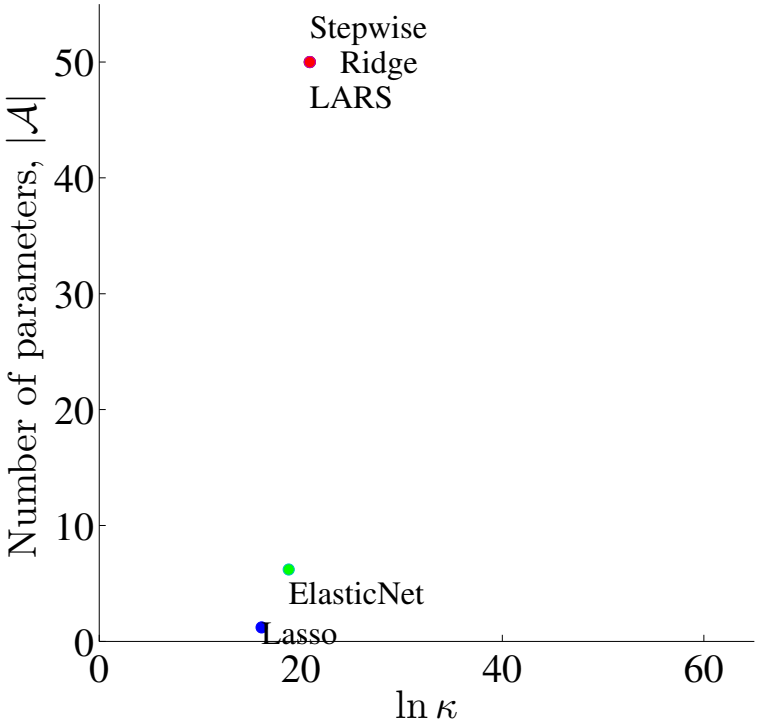


(a) $k = 0.2$

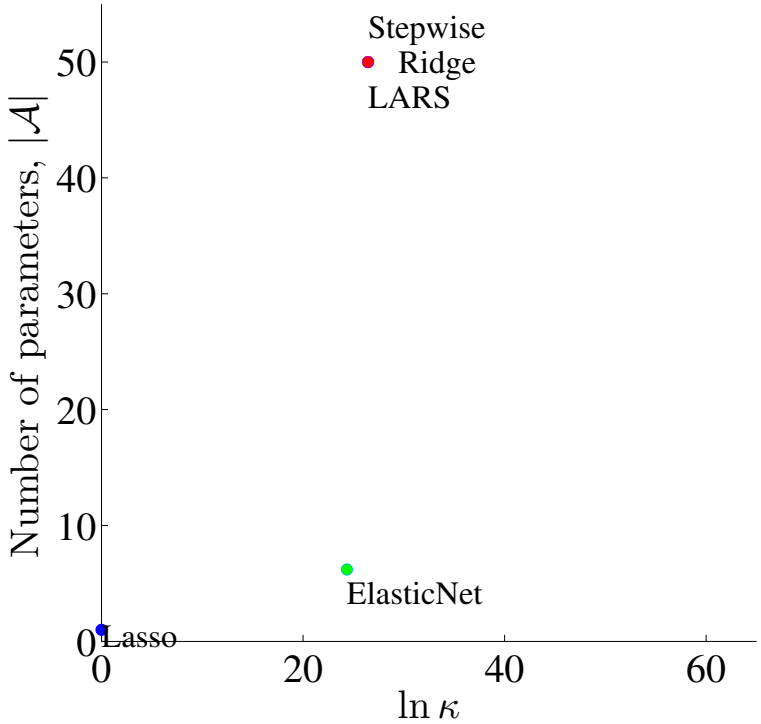


(b) $k = 0.8$

Сложность и устойчивость моделей для адекватных избыточных выборок

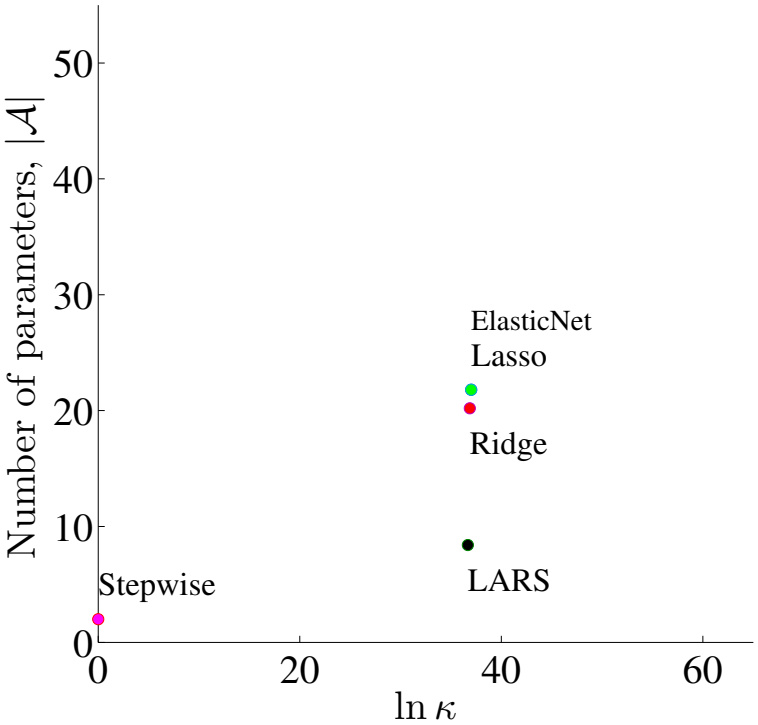


(a) $k = 0.2$

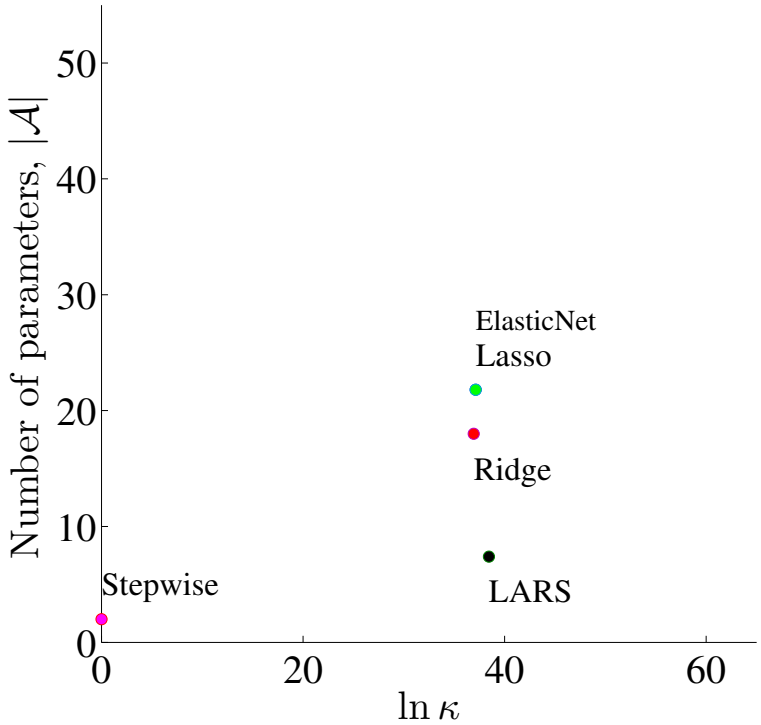


(b) $k = 0.8$

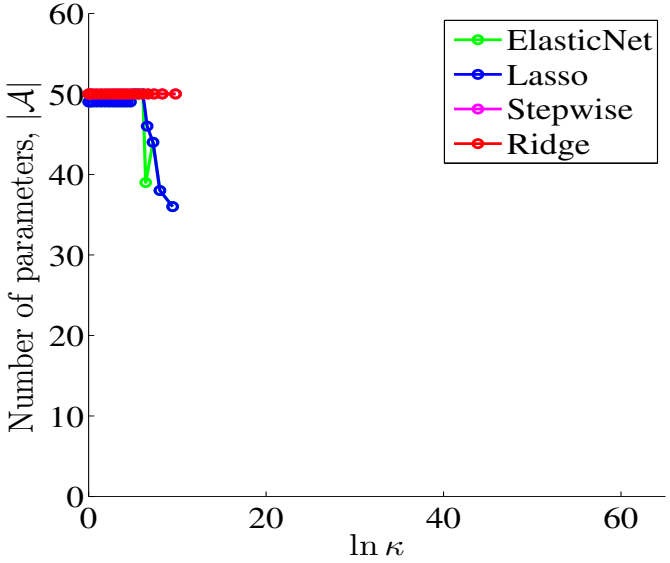
Сложность и устойчивость моделей для адекватных коррелирующих выборок



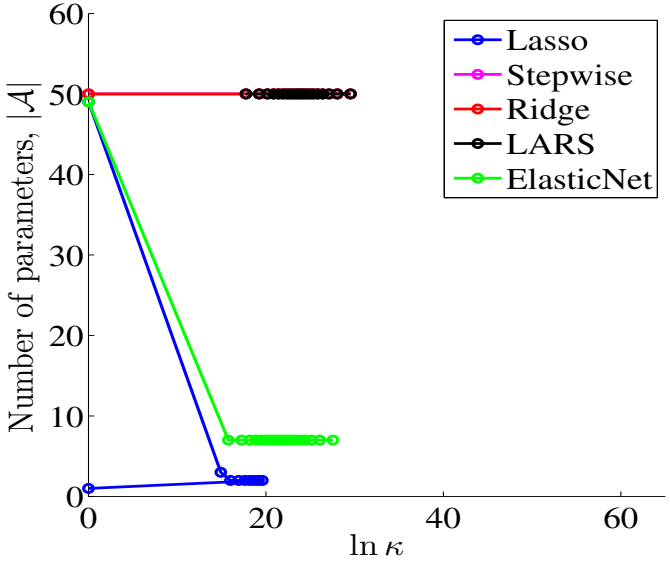
(a) $k = 0.2$



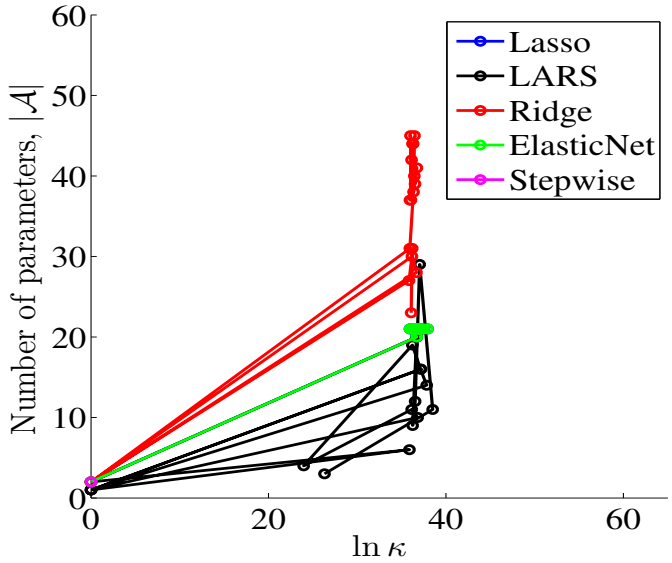
(b) $k = 0.8$



(a) Неадекватная коррелирующая



(b) Адекватная избыточная



(c) Адекватная кореллирующая

Заключение

- Предложен критерий ранжирования методов выбора признаков по количеству избыточных признаков в множестве отобранных признаков.
- Построена процедура генерации выборок для тестирования методов выбора признаков.
- Проведено ранжирование методов выбора признаков для сгенерированных тестовых выборок.

Публикации ВАК:

- Катруца А. М., Стрижов В. В. Проблема мультиколлинеарности при выбора признаков в регрессионных задачах // Информационные технологии. — 2014. — № 9. — ISSN 1684-6400 (принято в печать).
- A. Katrutsa, M. Kuznetsov, V. Strijov. Metric concentration search procedure using reduced matrix of pairwise distances // Intelligent Data Analysis. — 2014 (принято в печать).
- A. Katrutsa, V. Strijov. Stresstest procedures for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems (к подаче).