Выбор устойчивых прогностических моделей в задачах нелинейного регрессионного анализа

А. А. Токмакова

Научный руководитель: В. В. Стрижов
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

10 июня 2014 г.
Москва

Цель исследования

Предложить стратегию выбора устойчивых моделей в условиях мультиколлинеарности выборки в задачах нелинейного регрессионного анализа.

В работе предложена

Процедура получения модели с оптимальным числом параметров, использующая методы последовательного добавления и удаления элементов модели.

План презентации

1. Вид функции ошибки
2. Метод оценки ковариационной матрицы
3. Описание процедуры Add-Del

Постановка задачи

Выборка: $\mathfrak{D} = \{\mathbf{x}_i, y_i\}$, $i \in \mathcal{I} = \{1, \ldots, m\}$.

Регрессионная модель: $\mathbf{f} : (\mathbf{w}, \mathbf{X}) \mapsto \mathbf{y}$, где $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathbf{w} = [w_1, \ldots, w_j, \ldots, w_t]$, $j \in \mathcal{J} = \{1, \ldots, t\}$.

Нейронная сеть: $\mathbf{f}(\mathbf{x}, \mathbf{w}) = \mathbf{W}^{(2)}\mathbf{a}$, $\mathbf{a} = \mathbf{tanh}(\mathbf{W}^{(1)}\mathbf{x})$, где $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$ — матрицы весов первого и второго слоёв нейронной сети, $\mathbf{w} = \mathrm{vec}(\mathbf{W}^{(1)}|\mathbf{W}^{(2)})$.

Требуется найти такое множество индексов $\mathcal{A}^* \subseteq \mathcal{J}$, что:

$$\mathcal{A}^* = \operatorname*{argmin}_{\mathcal{A} \subseteq \mathcal{J}} S(\mathbf{w}^*_{\mathcal{A}} | \mathfrak{D}, \mathbf{f}), \quad \mathbf{w}^*_{\mathcal{A}} = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^t} S(\mathbf{w} | \mathfrak{D}, \mathbf{f}),$$

где $S(\mathbf{w})$ — функция ошибки модели.

Пусть многомерная случайная величина $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{B})$ имеет нормальное распределение

$$p(\mathbf{y}|\mathbf{X}, \mathbf{B}) = \frac{1}{(2\pi)^{\frac{m}{2}} \det^{\frac{1}{2}}(\mathbf{B}^{-1})} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^{\mathsf{T}}\mathbf{B}(\mathbf{y}-\mathbf{f})\right) = \frac{\exp(-E_D)}{Z_D(\mathbf{B})}.$$

Вектор параметров $\mathbf{w}$ модели $\mathbf{f}$ — многомерная случайная величина с математическим ожиданием $\mathbf{w}_0$ и ковариационной матрицей $\mathbf{A}^{-1}$

$$p(\mathbf{w}|\mathbf{A}) = \frac{1}{(2\pi)^{\frac{t}{2}} \det^{\frac{1}{2}}(\mathbf{A}^{-1})} \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\mathsf{T}}\mathbf{A}(\mathbf{w}-\mathbf{w}_0)\right) = \frac{\exp(-E_{\mathbf{w}})}{Z_{\mathbf{w}}(\mathbf{A})}.$$

Функция ошибки

$$p(\mathfrak{D}|\mathbf{w},\mathbf{B}) = \frac{\exp(-E_D)}{Z_D(\mathbf{B})},\ p(\mathbf{w}|\mathbf{A}) = \frac{\exp(-E_\mathbf{w})}{Z_\mathbf{w}(\mathbf{A})}.$$

Апостериорное распределение параметров модели для заданных ковариационных матриц $\mathbf{A}^{-1}$ и $\mathbf{B}^{-1}$ имеет вид:

$$p(\mathbf{w}|\mathfrak{D},\mathbf{A},\mathbf{B}) = \frac{p(\mathfrak{D}|\mathbf{w},\mathbf{B})p(\mathbf{w}|\mathbf{A})}{p(\mathfrak{D}|\mathbf{A},\mathbf{B})} = \frac{\exp\big(-(E_D+E_\mathbf{w})\big)}{Z_D Z_\mathbf{w}\ p(\mathfrak{D}|\mathbf{A},\mathbf{B})}.$$

Определим функцию ошибки $S(\mathbf{w})$:

$$S(\mathbf{w}) = E_\mathbf{w} + E_D = \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^\mathsf{T}\mathbf{A}(\mathbf{w}-\mathbf{w}_0) + \frac{1}{2}(\mathbf{y}-\mathbf{f})^\mathsf{T}\mathbf{B}(\mathbf{y}-\mathbf{f}).$$

Оценка ковариационной матрицы

Ковариационная матрица случайной величины $\mathbf{w} \in \mathbb{R}^t$:

$$\mathbf{A}^{-1} = \mathbb{E}\left[(\mathbf{w} - \mathbb{E}\mathbf{w})(\mathbf{w} - \mathbb{E}\mathbf{w})^\mathsf{T}\right] = \mathbb{E}[\mathbf{w}\mathbf{w}^\mathsf{T}] - \mathbb{E}[\mathbf{w}] \cdot \mathbb{E}[\mathbf{w}^\mathsf{T}].$$

Разобьем множество индексов выборки $\mathcal{I}$ на b подмножеств, на каждой подвыборке оценим вектор параметров $\mathbf{w}$ модели $\mathbf{f}$:

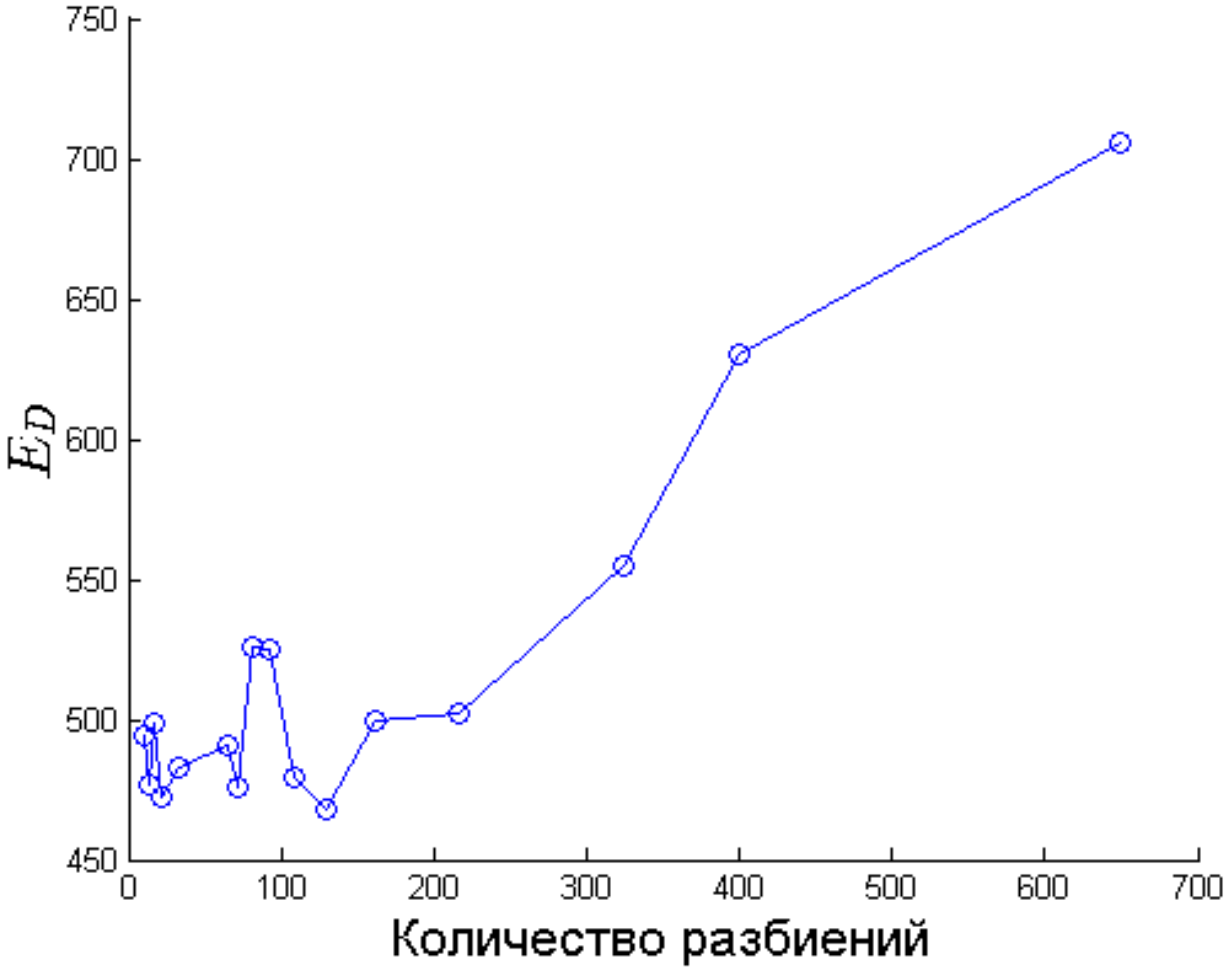
$$\mathbf{W} = [\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_b] \in \mathbb{R}^{t \times b}.$$

Центрируем каждую строку $\mathbf{W}_j - \mathbb{E}\mathbf{W}_j \mapsto \mathbf{W}_j$ и оценим ковариационную матрицу параметров модели $\mathbf{f}$:

$$\mathbf{A}^{-1} = \frac{1}{b}\mathbf{W}\mathbf{W}^\mathsf{T}.$$

Выбор величины b

Зависимость функции E_D от b (данные winequality, UCI):



Структура модели

Определим структуру модели как ациклический направленный граф с тремя множествами вершин фиксированной максимальной валентности:

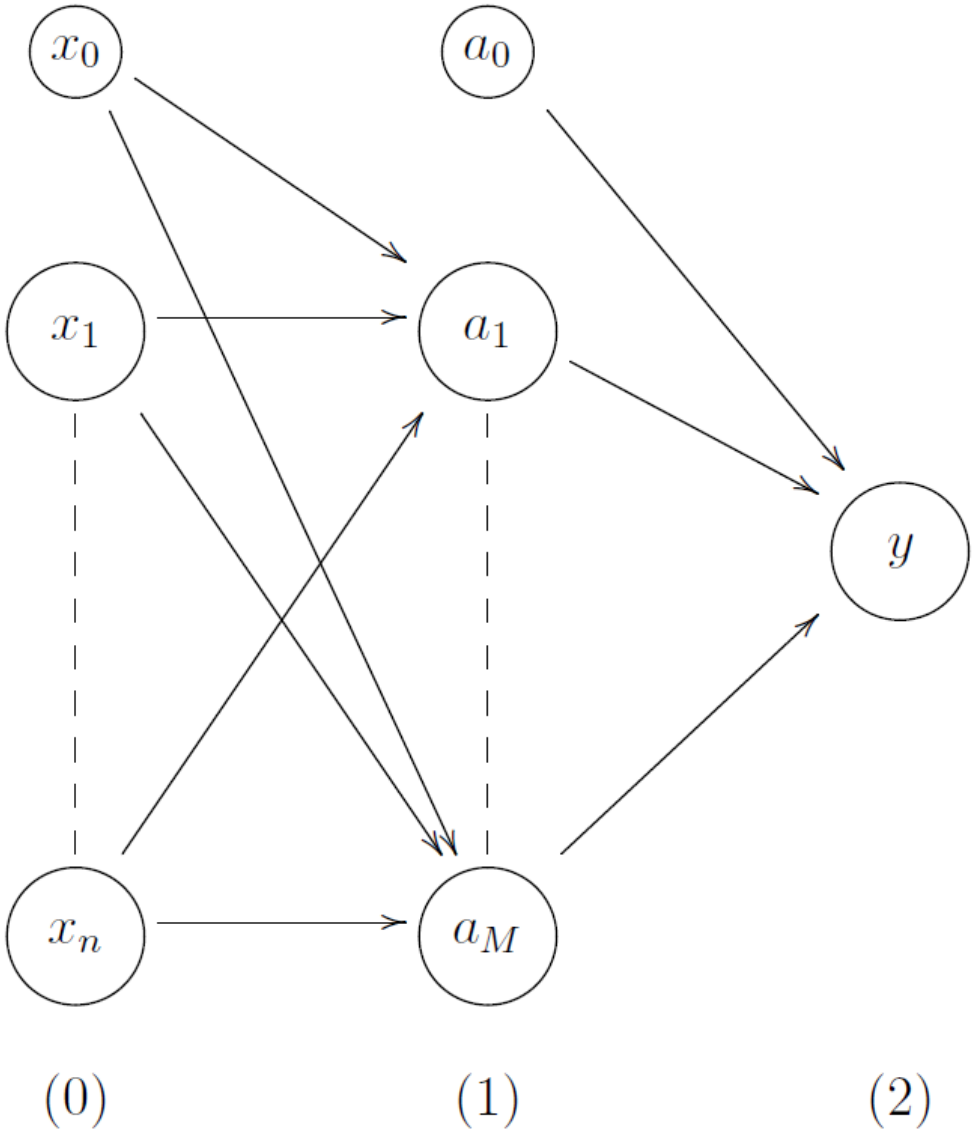
(0) — множество вершин свободных переменных размера $n + 1$;

(1) — множество вершин функции активации первого слоя размера $M + 1$;

(2) — множество вершин функции активации выходного слоя размера 1.

Ребра графа могут идти только из $(0) \to (1)$ или $(1) \to (2)$.

Представление нейронной сети в виде графа:



Последовательная модификация модели

Зададим вектор $\mathbf{z} \in \{0,1\}^t$:

$$z_j = \begin{cases} 0, & \text{если } w_j + \Delta w_j = 0, \text{ т. е. ребра нет в графе;} \\ 1, & \text{иначе.} \end{cases}$$

$\mathbf{z}$, n и M однозначно задают вид графа, а значит и структуру модели.

Представим процедуру последовательной модификации структуры модели в виде пути внутри t-мерного гиперкуба $\mathcal{Z}$, каждая вершина которого является бинарным вектором $\mathbf{z} = \{0,1\}^t$.

Соседние вершины гиперкуба $\mathcal{Z}$ отличаются единственной компонентой вектора $\mathbf{z}$.

Определение стратегии алгоритма Add-Del

Стратегия задается следующими математическими объектами:

- набором критериев качества $\{Q\}$;
- набором ограничений на структуру и параметры модели $\mathcal{A} \in \mathcal{J}$, $\mathbf{w} = \mathbf{w}^*_{\mathcal{A}}$;
- критериями останова шагов удаления и добавления структурных единиц в модель;
- критерием останова процедуры выбора модели.

То есть, действуя согласно стратегии, будем изменять структуру модели, удаляя из неё элементы и добавляя их до тех пор, пока значение критериев качества не стабилизируется.

Добавление структурных связей

Пусть на шаге q известна правдоподобная нелинейная модель субоптимальной сложности, то есть известны оценки её параметров $\mathbf{w}^*_{\mathcal{A}_q}$ и гиперпараметров $\mathbf{w}^*_{0q}$ и $\mathbf{A}^*_q$:

$$\mathbf{f}_{\mathcal{A}_q}\Big|_{\mathbf{w}=\mathbf{w}^*_{\mathcal{A}_q}} : \mathbf{X} \mapsto \mathbf{y}.$$

1. Найдем такой элемент множества $\mathcal{J} \setminus \mathcal{A}_q$, что:

$$j^* = \operatorname*{argmin}_{j \in \mathcal{J} \setminus \mathcal{A}_q} S(\mathbf{w}_{\mathcal{A}_q \cup \{j\}} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_q \cup \{j\}}).$$

2. Добавим новый элемент j^* к текущему набору

$$\mathcal{A}_{q+1} = \mathcal{A}_q \cup \{j^*\}$$

и будем повторять эту процедуру, пока

$$\left| S(\mathbf{w}^*_{\mathcal{A}_{q+1}} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_{q+1}}) - S(\mathbf{w}^*_{\mathcal{A}_q} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_q}) \right| \leq \Delta S_1.$$

Удаление структурных связей. Метод Белсли.

Представим оценку ковариационной матрицы на шаге q как $\mathbf{A}_q^{-1} = (\mathbf{L}\mathbf{L}^{\mathsf{T}})^{-1}$. Запишем сингулярное разложение матрицы $\mathbf{L}$:

$$\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}},$$

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

$$\lambda_1 > \lambda_2 > \ldots > \lambda_q > 0.$$

Индексом обусловленности с индексом j будем называть:

$$\eta_j = \frac{\lambda_{\max}}{\lambda_j}.$$

Запишем матрицу $\mathbf{A}_q^{-1}$ в виде:

$$\mathbf{A}_q^{-1} = (\mathbf{L}\mathbf{L}^{\mathsf{T}})^{-1} = (\mathbf{V}\mathbf{\Lambda}^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}})^{-1} = \mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^{-1}.$$

Оценками дисперсии параметров будут диагональные элементы матрицы $\mathbf{A}_q^{-1}$:

$$\alpha_{ii} = \sum_{j=1}^{t} \frac{v_{ij}^2}{\lambda_j^2}.$$

Дисперсионной долей r_{ij} будем называть вклад j-го признака в дисперсию i-го элемента вектора параметров $\mathbf{w}$:

$$r_{ij} = \frac{v_{ij}^2/\lambda_j^2}{\sum\limits_{j=1}^{t} v_{ij}^2/\lambda_j^2}.$$

Удаление структурных связей

1. Вычислим индексы обусловленности η_j и матрицу долевых коэффициентов $\mathfrak{R} = [r_{gj}]$ для $\mathbf{w}_{\mathcal{A}_q}$.
2. Найдем индекс g^* максимального индекса обусловленности $\eta_{\max}$.
3. В матрице долевых коэффициентов $\mathfrak{R}$ найдем индекс столбца j^*.

$$j^* = \operatorname*{argmax}_{j \in \mathcal{A}_q} s_{g^* j}.$$

4. Удалим индекс j^* из множества $\mathcal{A}_q$:

$$\mathcal{A}_{q+1} = \mathcal{A}_q \setminus \{j^*\}$$

и будем повторять эту процедуру, пока

$$\left| S(\mathbf{w}^*_{\mathcal{A}_{q+1}} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_{q+1}}) - S(\mathbf{w}^*_{\mathcal{A}_q} | \mathfrak{D}, \mathbf{f}_{\mathcal{A}_q}) \right| \leq \Delta S_2.$$

Выборка

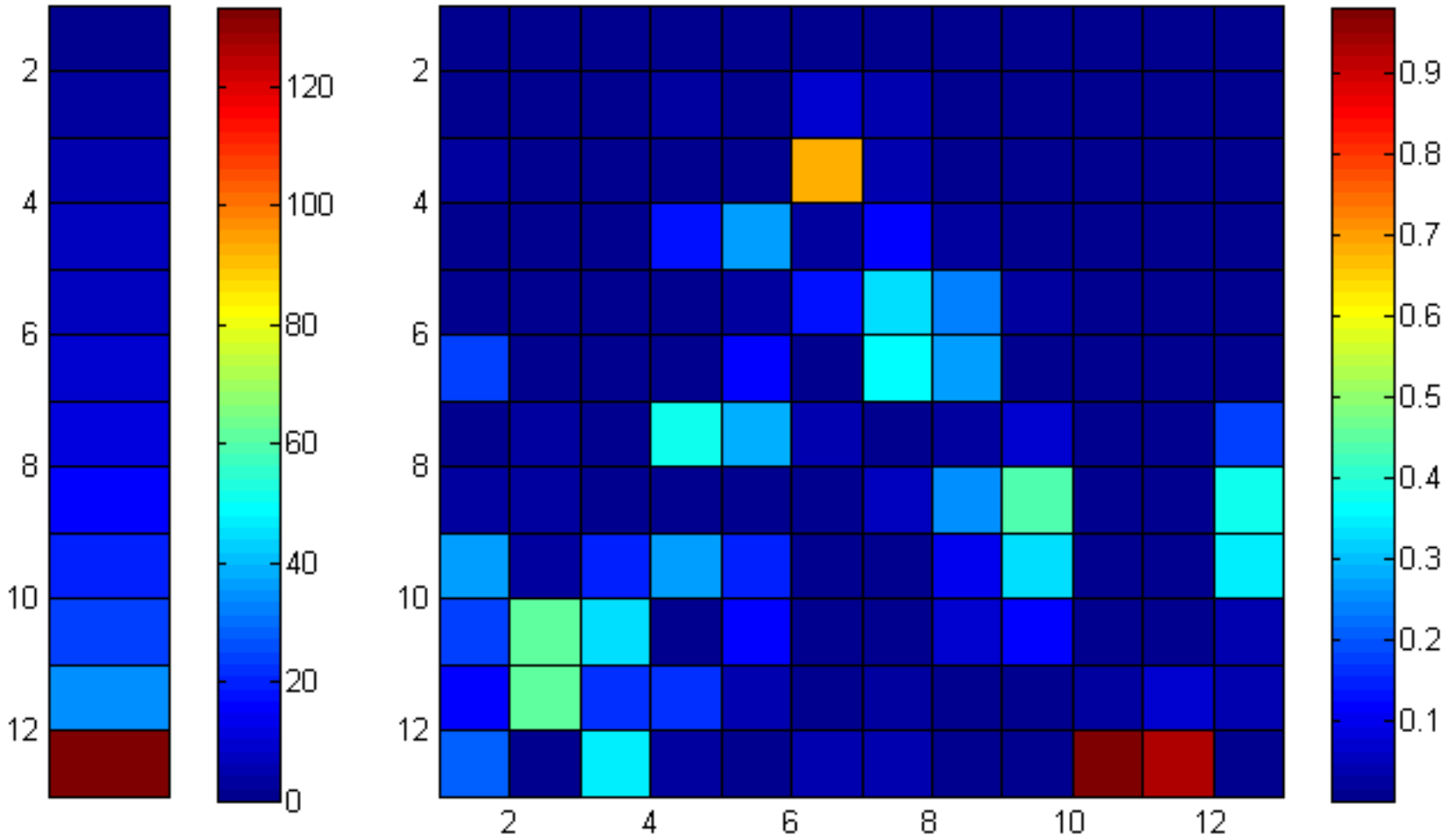


Рис.: индексы обусловленностей и матрица дисперсионных долей.

Критерий останова процедуры Add-Del

В качестве критерия стабилизации структуры модели предлагается использовать энтропию:

$$H(\mathcal{A}, \mathcal{A}') = -\rho(\mathbf{z}, \mathbf{z}') \ln\big(\rho(\mathbf{z}, \mathbf{z}')\big),$$

где $\rho(\cdot,\cdot)$ — расстояние Хэмминга между векторами $\mathbf{z}$ и $\mathbf{z}'$, где

$$z_j = \begin{cases} 0, & \text{если } w_j + \Delta w_j = 0, \text{ т. е. } j \notin \mathcal{A}; \\ 1, & \text{иначе.} \end{cases}$$

Процесс считается стабильным, если изменение энтропии $H(\mathcal{A}, \mathcal{A}')$ не превосходит заданного порога.

Подбор начального приближения

Рассмотрим последовательно порожденные модели $\mathbf{f}_{\mathcal{A}_q}$ и $\mathbf{f}_{\mathcal{A}_{q+1}}$. Начальные приближения гиперпараметров для модели $\mathbf{f}_{\mathcal{A}_{q+1}}$ будут выглядеть следующим образом:

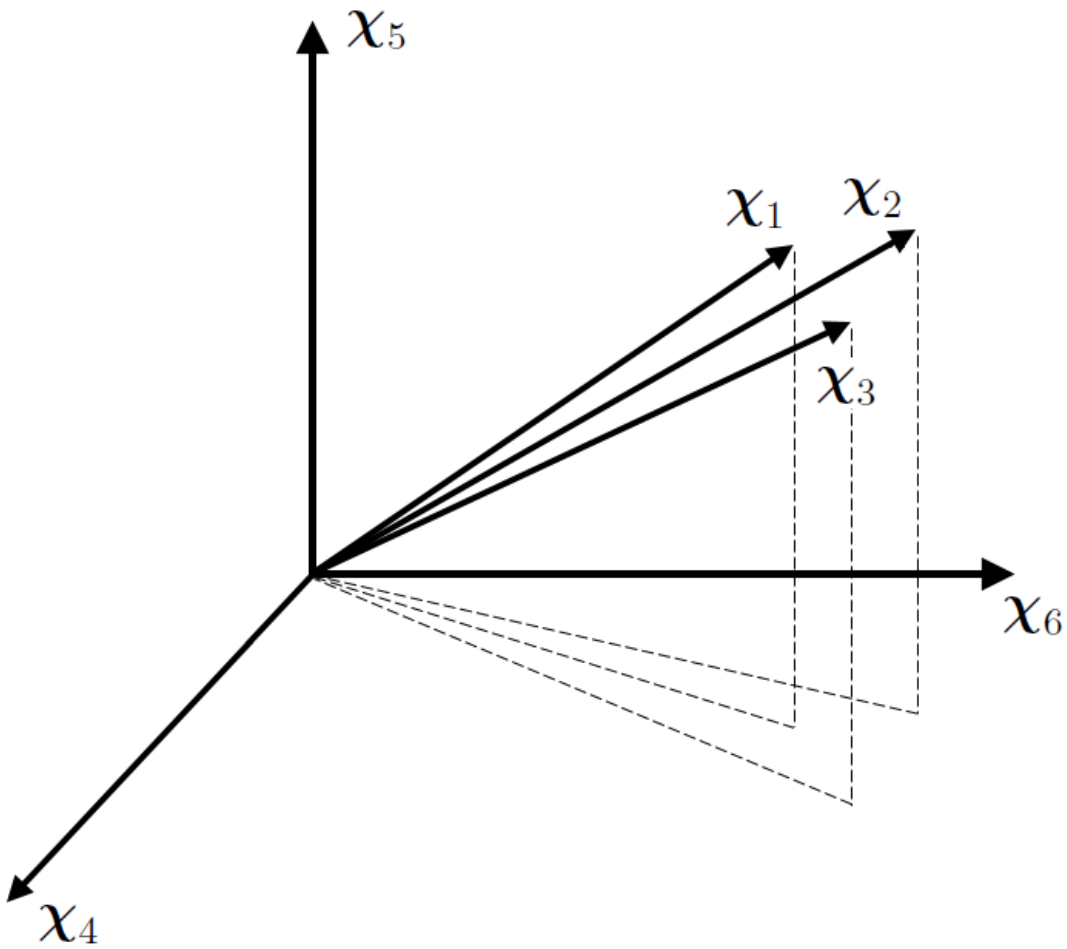
$$\tilde{\mathbf{w}}_{\mathcal{A}_q} = [\mathbf{w}_{\mathcal{A}_q}; 0]^{\mathsf{T}}; \ \tilde{\mathbf{A}}_q = \begin{pmatrix} \mathbf{A}_q & 0 \\ 0 & 1 \end{pmatrix}.$$

Запишем функцию правдоподобия данных и априорное распределение параметров модели $\mathbf{f}_{\mathcal{A}_{q+1}}$ следующим образом:

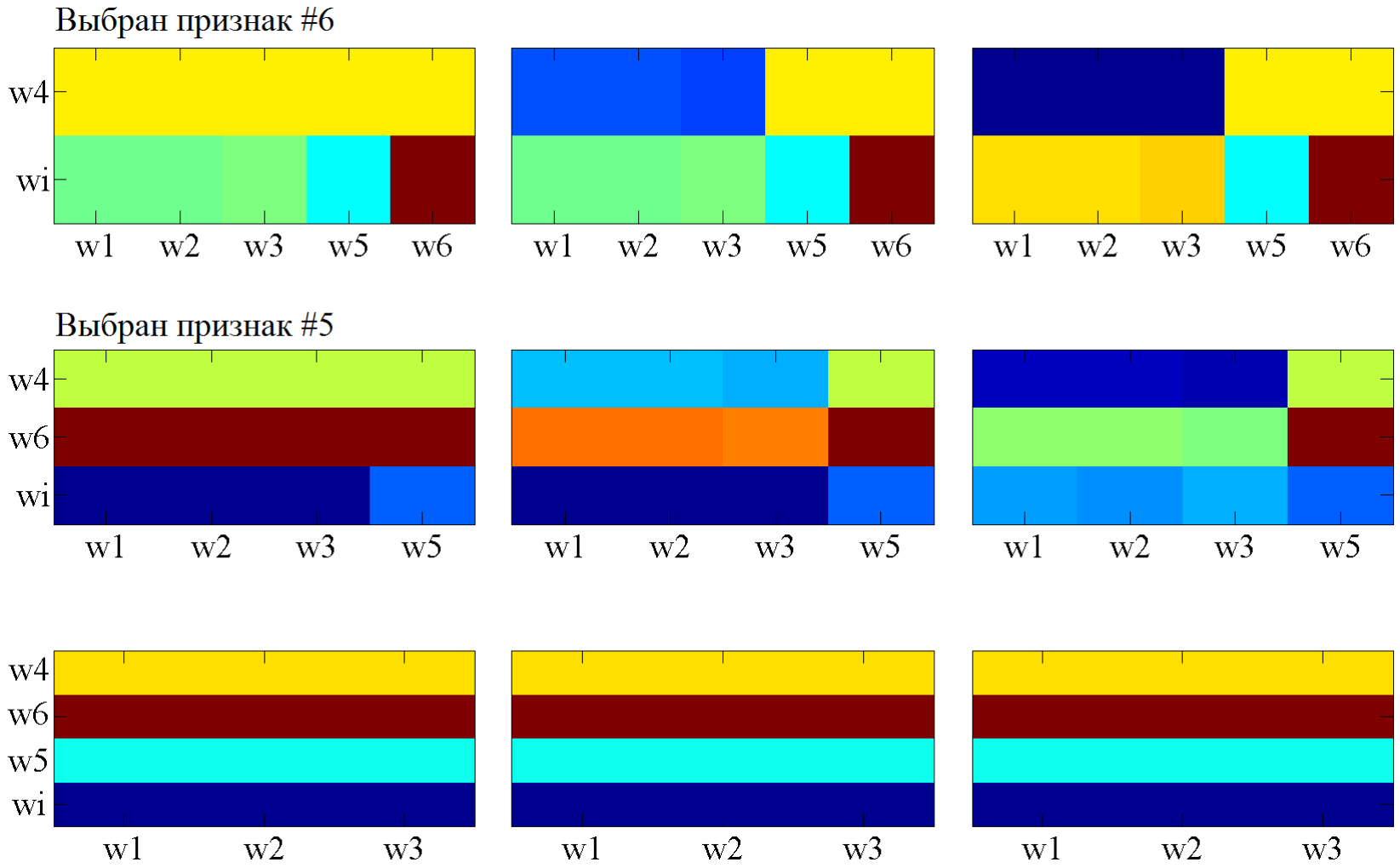
$$p(\mathfrak{D}|\mathbf{w}_{\mathcal{A}_{q+1}}, \mathbf{B}_q) = \frac{1}{(2\pi)^{\frac{m}{2}} \det^{\frac{1}{2}}(\mathbf{B}_q^{-1})} \cdot \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{f}_{\mathcal{A}_{q+1}})^{\mathsf{T}} \mathbf{B}_q (\mathbf{y} - \mathbf{f}_{\mathcal{A}_{q+1}})\right);$$

$$p(\mathbf{w}_{\mathcal{A}_{q+1}}|\tilde{\mathbf{A}}_q) = \frac{1}{(2\pi)^{\frac{t}{2}} \det^{\frac{1}{2}}(\tilde{\mathbf{A}}_q^{-1})} \cdot \exp\left(-\frac{1}{2}(\mathbf{w}_{\mathcal{A}_{q+1}} - \tilde{\mathbf{w}}_{\mathcal{A}_q})^{\mathsf{T}} \tilde{\mathbf{A}}_q (\mathbf{w}_{\mathcal{A}_{q+1}} - \tilde{\mathbf{w}}_{\mathcal{A}_q})\right).$$

$$\mathbf{y} = 0.3\boldsymbol{\chi}_5 + 0.3\boldsymbol{\chi}_6 + 0.3\boldsymbol{\chi}_3 + \mathcal{U}[0,1].$$



$$\mathbf{y} = 0.3\boldsymbol{\chi}_5 + 0.3\boldsymbol{\chi}_6 + 0.3\boldsymbol{\chi}_3 + \mathcal{U}[0,1],\ \hat{\mathbf{y}} = 0.3\boldsymbol{\chi}_5 + 0.8\boldsymbol{\chi}_6 + 0.5\boldsymbol{\chi}_4.$$



Алгоритм Левенберга-Марквардта

Алгоритм Левенберга-Марквардта предназначен для оптимизации параметров нелинейных регрессионных моделей.

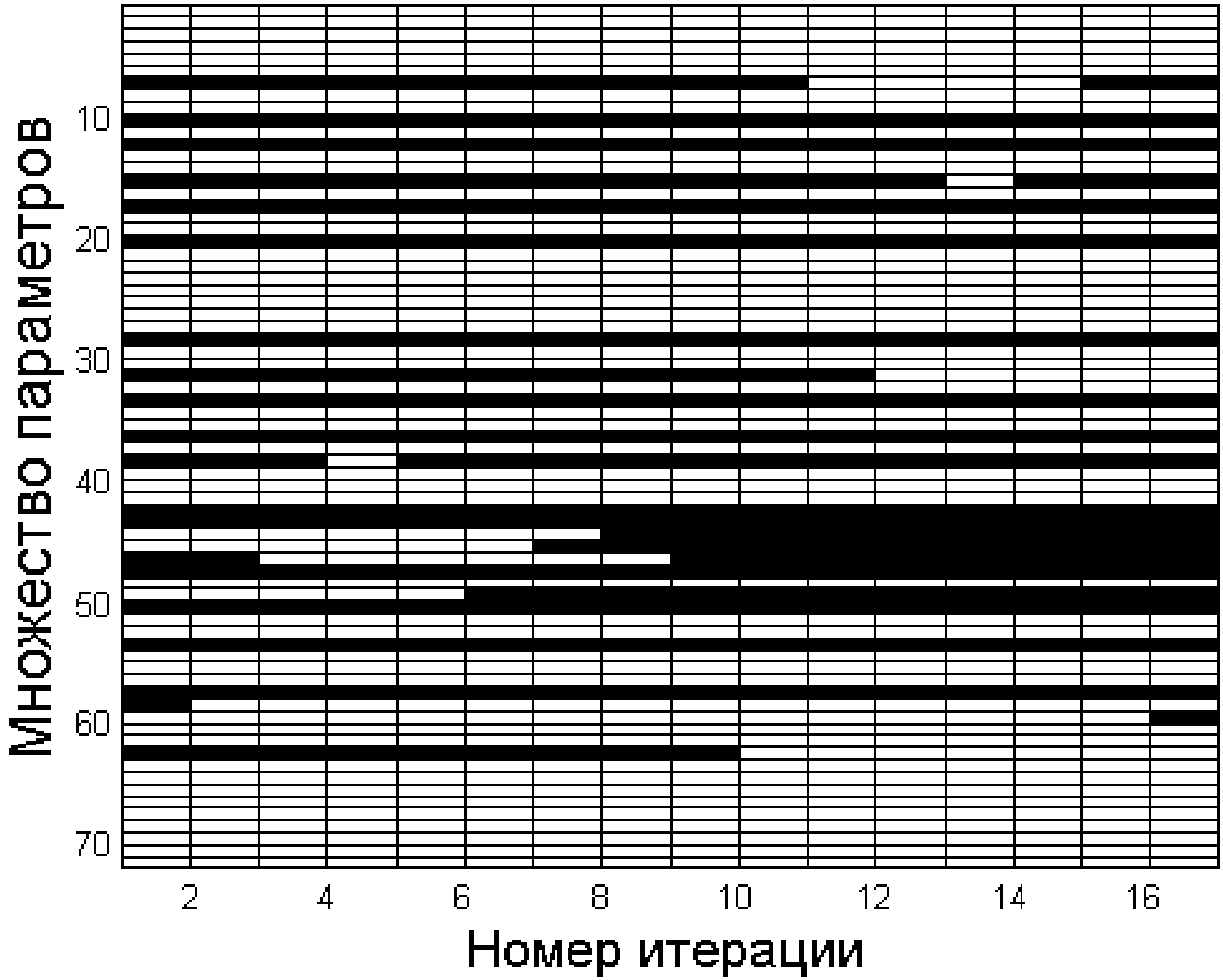
$\mathbf{f}(\mathbf{w}+\Delta\mathbf{w},\mathbf{X}) \approx \mathbf{f}(\mathbf{w},\mathbf{X}) + \mathbf{J}\Delta\mathbf{w}$, где $\mathbf{J}$ — якобиан функции $\mathbf{f}$ в точке $\mathbf{w}$.

Функция ошибки:
$S = \frac{1}{2}(\mathbf{w}+\Delta\mathbf{w})^T\mathbf{A}(\mathbf{w}+\Delta\mathbf{w}) + \frac{1}{2}\left(\mathbf{f}(\mathbf{w}+\Delta\mathbf{w},\mathbf{X}) - \mathbf{y}\right)^T\mathbf{B}\left(\mathbf{f}-\mathbf{y}\right).$

Для нахождения экстремума приравняем вектор частных производных S по $\mathbf{w}$ к нулю: $\nabla S = \frac{1}{2}(\mathbf{w}+\Delta\mathbf{w})^T(\mathbf{A}+\mathbf{A}^T) + \frac{1}{2}\left[\left(\mathbf{J}(\mathbf{w}+\Delta\mathbf{w}) - \mathbf{y}\right)^T\mathbf{B}^T\mathbf{J} + \left(\mathbf{J}(\mathbf{w}+\Delta\mathbf{w}) - \mathbf{y}\right)^T\mathbf{B}\mathbf{J}\right] = 0.$

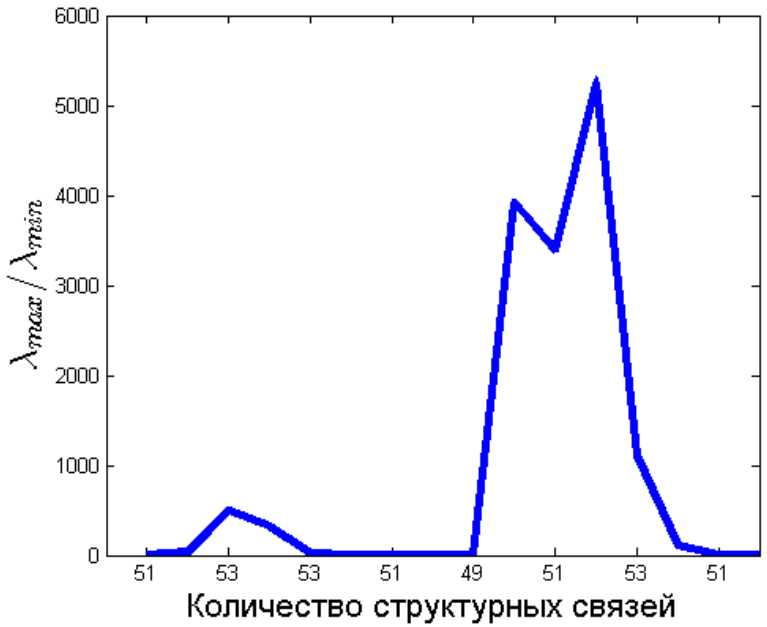
Выразив приращение $\Delta\mathbf{w}$, получим следующую рекуррентную формулу: $\Delta\mathbf{w} = \left[\left(\mathbf{A}+\mathbf{A}^T + \mathbf{J}^T(\mathbf{B}^T+\mathbf{B})\mathbf{J}\right)^{-1}\right]^T\left(-\mathbf{w}^T(\mathbf{A}+\mathbf{A}^T) + (\mathbf{y}-\mathbf{f}(\mathbf{w},\mathbf{X}))^T(\mathbf{B}^T+\mathbf{B})\mathbf{J}\right)^T.$

Результаты


Множество параметров
10
20
30
40
50
60
70
2
4
6
8
10
12
14
16
Номер итерации

Результаты


600
550
500
450
400
350
300
250
200
150
100
E_D
51 53 53 51 49 51 53 51
Количество структурных связей


6000
5000
4000
3000
2000
1000
0
$\lambda_{max} / \lambda_{min}$
51 53 53 51 49 51 53 51
Количество структурных связей

Предложено

- стратегия пошаговой модификации структуры модели;
- модификация алгоритма Левенберга-Марквардта для градиентной оптимизации в случае нелинейных моделей.

Публикации ВАК

1. Токмакова А.А., Стрижов В.В. *Оценивание гиперпараметров линейных регрессионных моделей при отборе шумовых и коррелирующих признаков*. — Информатика и её применения, 2012. Том 6(4), сс. 66-75.
2. Оценка гиперпараметров регрессионных моделей методом максимального правдоподобия. — Информационные технологии, 2013. Том 2, сс. 11-15.
3. Токмакова А.А. *Алгоритм стохастического отбора объектов и признаков в задаче банковского кредитного скоринга*. — Информационные технологии, 2014. Том 3, сс. 30-35.
4. Kuznetsov M.P., Tokmakova A.A., Strijov V.V. *Analytic and stochastic methods of structure parameter estimation*. — Machine Learning Journal, 2014.