Оценивание гиперпараметров линейных регрессионных моделей при отборе шумовых и коррелирующих признаков

А. А. Токмакова

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

13 июня 2012 г.
Москва

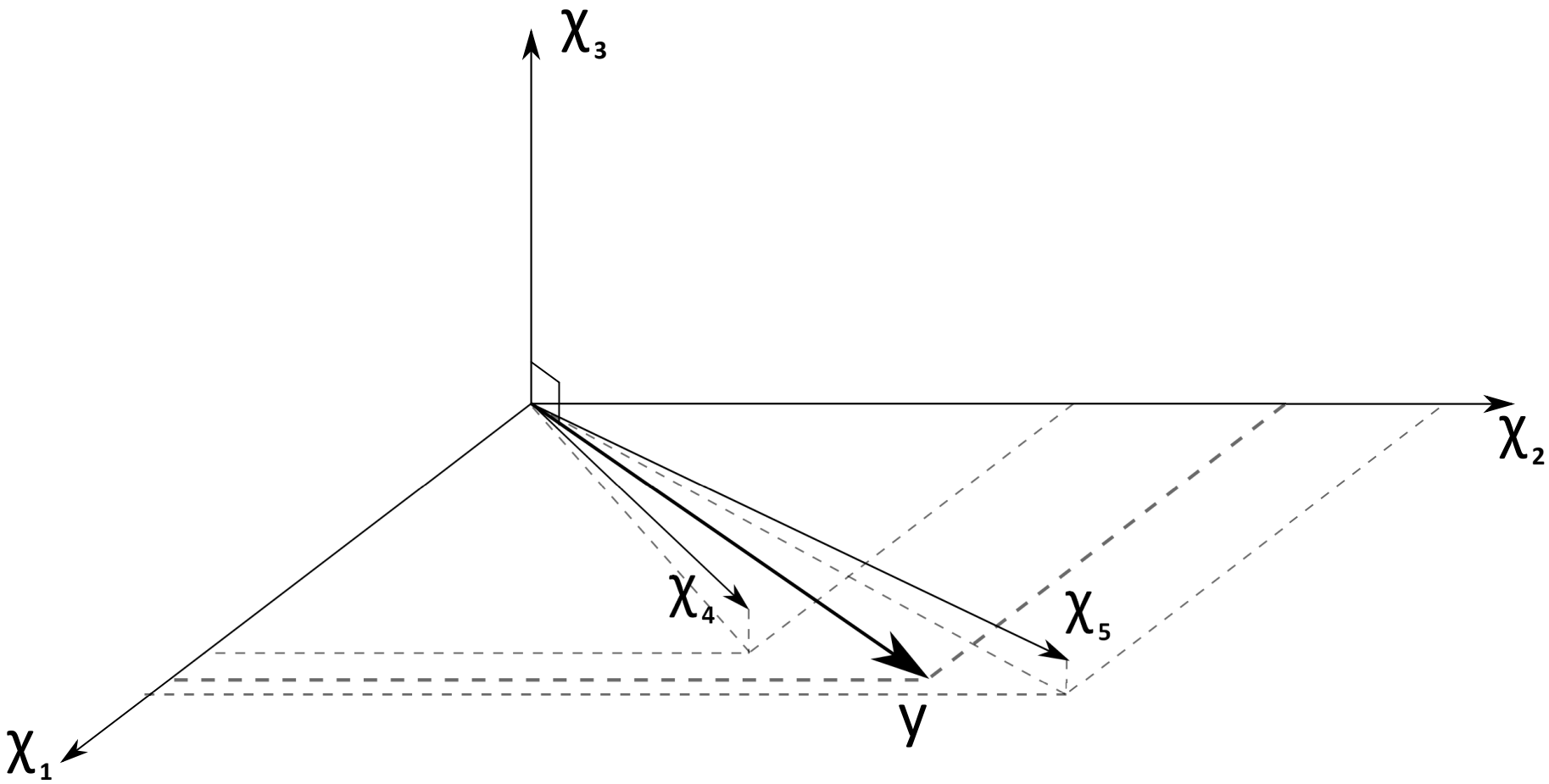
Цель исследования

Необходимо произвести отбор шумовых и коррелирующих признаков линейной модели, а также оценить ковариационную матрицу параметров модели.

План презентации

1. Постановка задачи
2. Функция ошибки
3. Аппроксимация Лапласа
4. Оценка ковариационных матриц
5. Модифицированный алгоритм Левенберга-Марквардта
6. Вычислительный эксперимент
7. Результаты

Шумовые и мультикоррелирующие признаки



Здесь χ_4, χ_5 — коррелирующие признаки, а χ_3 — шумовой.
Множество индексов активных признаков $\mathcal{A} = \{1; 2\}$.

Регрессионная выборка: $D = \{\mathbf{x}_i, y_i\}_{i=1}^m = (X, \mathbf{y})$,
где $\mathbf{x}_i \in \mathbb{R}^n, i = 1, \ldots, m$.

Модель: $\mathbf{f}(\mathbf{w}, X) = X\mathbf{w}$.

Зависимая переменная: $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, B^{-1})$.

Задача: найти $\mathcal{A}^*$, которое бы доставляло минимум функции:

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

где $S(\mathbf{f}|\mathbf{w}, D)$ — функция ошибки.

При этом:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w}|\mathbf{f}_\mathcal{A}, D).$$

Функция ошибки в общем виде

Введём обозначние: $p(\mathbf{y}|X, \mathbf{w}, B, \mathbf{f}) \stackrel{\text{def}}{=} p(D|\mathbf{w}, B, \mathbf{f})$.

Запишем плотность $\mathbf{y}$ в виде:

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

Так как $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, B^{-1})$ и используется линейная модель, значит что $\mathbf{w} \sim \mathcal{N}(0, A^{-1})$.

$$p(\mathbf{w}|A, \mathbf{f}) = \frac{1}{(2\pi)^{\frac{n}{2}}|A|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{w}^T A\mathbf{w}\right) = \frac{\exp(-E_{\mathbf{w}})}{Z_{\mathbf{w}}(A)}.$$

Функция ошибки в общем виде

Запишем формулу Байеса для апостериорного распределения параметров модели:

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

Записывая функцию ошибки как:

$$S(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T A\mathbf{w} + \frac{1}{2}(\mathbf{y} - \mathbf{f})^T B(\mathbf{y} - \mathbf{f}),$$

получим следующее выражение для апостериорного распределения параметров:

$$p(\mathbf{w}|D, A, B, \mathbf{f}) = \frac{\exp\big(-S(\mathbf{w})\big)}{Z_S(A, B)},$$

где оценка нормировочного коэффициента $Z_S(A, B)$ производится с помощью аппроксимации Лапласа.

Аппроксимация Лапласа

Ненормированное распределение $p^*(\mathbf{w}) = \exp\bigl(-S(\mathbf{w})\bigr)$.

Нормировочная константа: $Z_S = \int p^*(\mathbf{w})d\mathbf{w}$ — ?

1. $\ln p^*(\mathbf{w}) = -S(\mathbf{w}) \approx -S(\mathbf{w}_0) + 0 - \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T H(\mathbf{w}-\mathbf{w}_0)$, где $H = -\nabla^2 S(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_0}$ — матрица Гессе.
2. $p^*(\mathbf{w}) \approx p^*(\mathbf{w}_0)\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T H(\mathbf{w}-\mathbf{w}_0)\right)$.
3. $\widehat{p}(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, H^{-1}) =$

   $= \frac{1}{(2\pi)^{\frac{n}{2}}|H|^{-\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T H(\mathbf{w}-\mathbf{w}_0)\right)$.

Нормировочная константа: $Z_S = \exp\bigl(-S(\mathbf{w}_0)\bigr)\frac{(2\pi)^{\frac{n}{2}}}{|H|^{\frac{1}{2}}}$.

Максимизация правдоподобия модели

Для нахождения гиперпараметров воспользуемся принципом максимума правдоподобия $p(D|A,B)$ относительно A и B. Запишем $p(D|A,B)$ в следующем виде:

$$p(D|A,B,\mathbf{f}) = \int p(D|\mathbf{w},B)p(\mathbf{w}|A)d\mathbf{w} = \frac{\int \exp(-S(\mathbf{w}))d\mathbf{w}}{Z_{\mathbf{w}}(A)Z_D(B)} \to \max.$$

Из условия нормировки:

$$\int p(\mathbf{w}|D,A,B)d\mathbf{w} = \frac{\int \exp(-S(\mathbf{w}))d\mathbf{w}}{Z_S(A,B)} = 1.$$

Таким образом, получим оценку логарифма правдоподобия:

$$\ln p(D|A,B,\mathbf{f}) = \frac{1}{2}\ln|A| - \frac{m}{2}\ln 2\pi + \frac{1}{2}\ln|B| - S(\mathbf{w}_0) - \frac{1}{2}\ln|H|.$$

Оценка ковариационных матриц

Рассмотрим случай, когда матрица $A = \text{diag}(\alpha_1, \ldots, \alpha_n)$ диагональна, а $B = \beta I_m$. Представим гессиан H в виде:

$$H = -\nabla\nabla S(\mathbf{w}) = -\nabla\nabla(\beta E_D + E_{\mathbf{w}}) = -\beta\nabla\nabla E_D - \nabla\nabla E_{\mathbf{w}} = H_D + H_{\mathbf{w}}.$$

Так как $\nabla\nabla E_{w_i} = \nabla\nabla(\frac{1}{2}\alpha_i(w_i - w_{0i})^2) = \alpha_i$, следовательно $H_{\mathbf{w}}$ — диагональная матрица.

H_D также диагональная матрица. Для этого рассмотрим два случая:

1. если все признаки независимы;
2. в выборке присутствуют шумовые или коррелирующие признаки.

Таким образом представим H_D как: $H_D = \beta\text{diag}(h_1, ..., h_n)$.

Частная производная по α:

$$\frac{1}{\alpha_i} - (w_i - w_0)^2 - \frac{1}{\beta h_i + \alpha_i} = 0.$$

По критерию Сильвестра матрица A не имеет отрицательных компонент:

$$\alpha_i = \frac{1}{2}\lambda_i\left(\sqrt{1 + \frac{4}{(w_i - w_0)^2\lambda_i}} - 1\right),$$

где $\lambda_i = \beta h_i$.

Частная производная по β:

$$\frac{m}{2\beta} - E_D - \frac{1}{2\beta}\gamma = 0; \quad \gamma = \sum_{j=1}^{W} \frac{\lambda_j}{\lambda_j + \alpha_j}.$$

Таким образом

$$\beta = \frac{m - \gamma}{2E_D}.$$

Оценка ковариационных матриц

Итерационный процесс:

1. вычисляем $\mathbf{w}$
   $S(\mathbf{w}) \to \min\limits_{\mathbf{w}}$;
2. пересчитываем гиперпараметры α и β.

При наличии шумовых и коррелирующих признаков необходимо принудительно занижать возрастающие диагональные элементы.

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

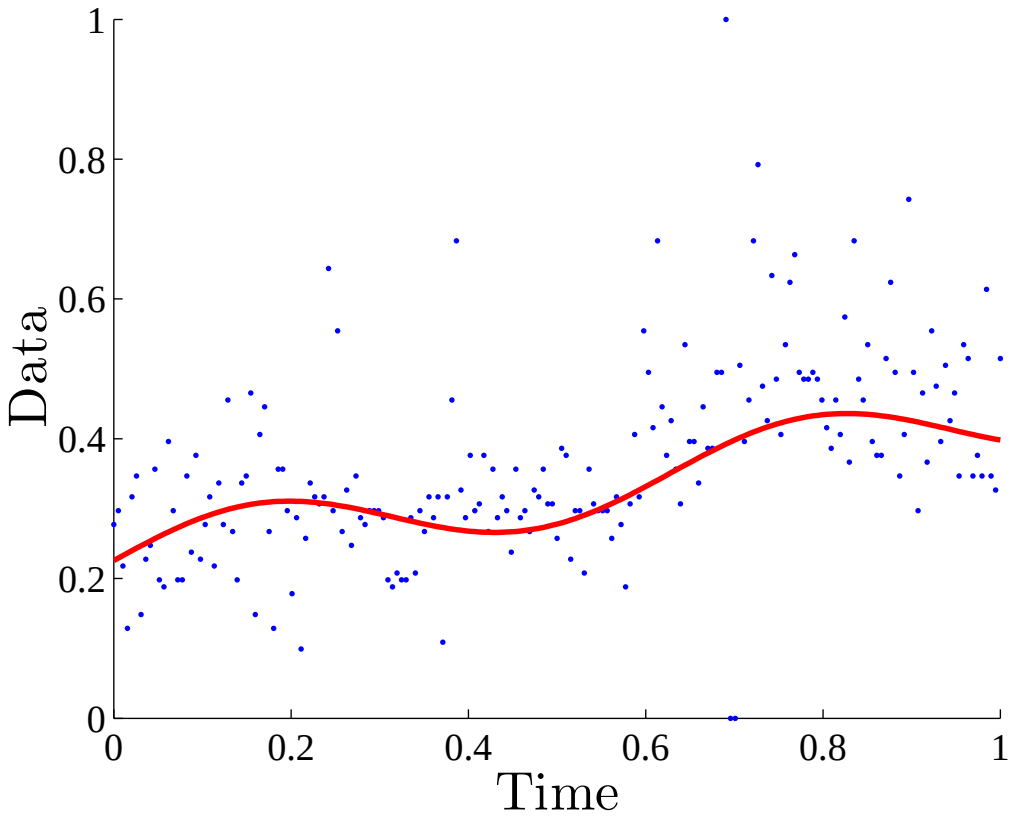
Функция ошибки:
$S = (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f})$.

Формула для приращения $\Delta\mathbf{w}$:
$\Delta\mathbf{w} = (J^T J)^{-1} J^T(\mathbf{y} - \mathbf{f})$.

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

Формула для приращения $\Delta\mathbf{w}$:
$\Delta\mathbf{w} = \left[\left(A + A^T + J^T(B^T + B)J\right)^{-1}\right]^T\left(-\mathbf{w}^T(A + A^T) + (\mathbf{y} - \mathbf{f}(\mathbf{w}, X))^T(B^T + B)J\right)^T$.

Вычислительный эксперимент

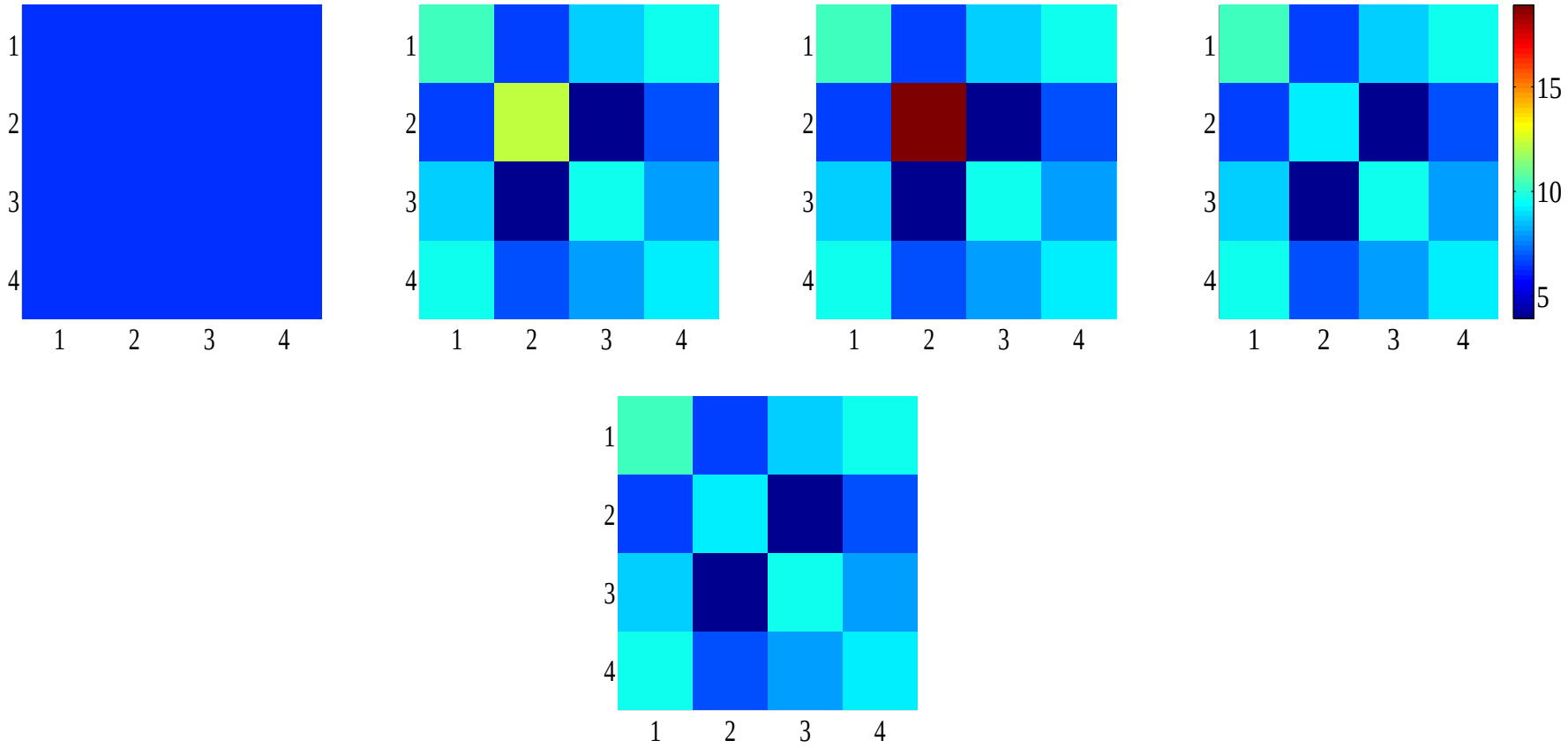


$$\mathbf{y} = 0.2256 + 0.1996\boldsymbol{\xi} + 0.0496\sin(10\boldsymbol{\xi}).$$

$$\boldsymbol{\chi}_1 = \boldsymbol{\xi}^0; \boldsymbol{\chi}_2 = \boldsymbol{\xi}^1; \boldsymbol{\chi}_3 = \sin(10\boldsymbol{\xi}).$$

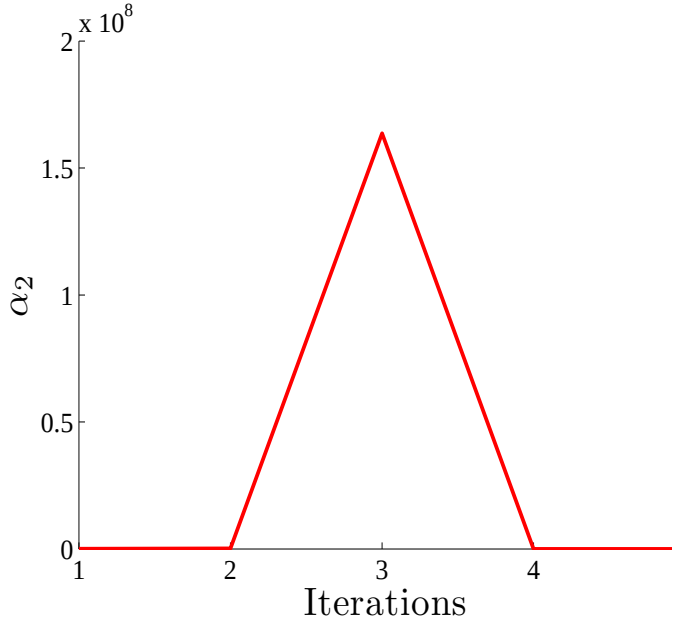
Случай шумовых признаков

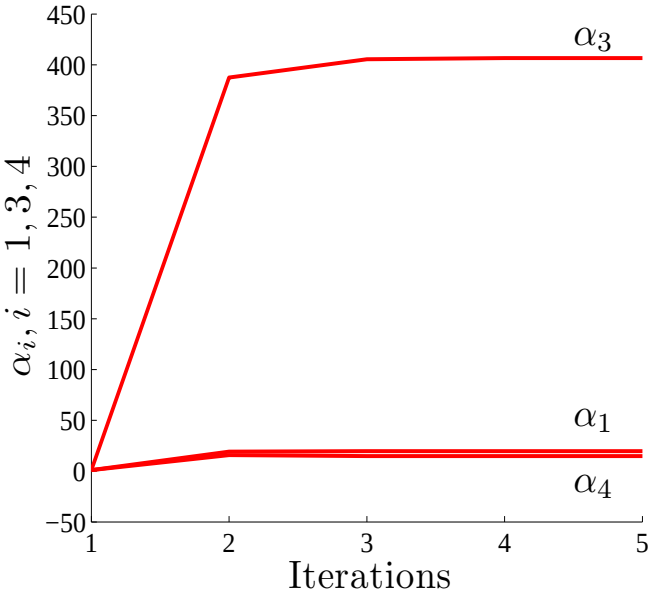
$$\chi_1 = \xi^0;\ \chi_2 \sim \mathcal{N}(0,1);\ \chi_3 = \xi^1;\ \chi_4 = \sin(10\xi).$$

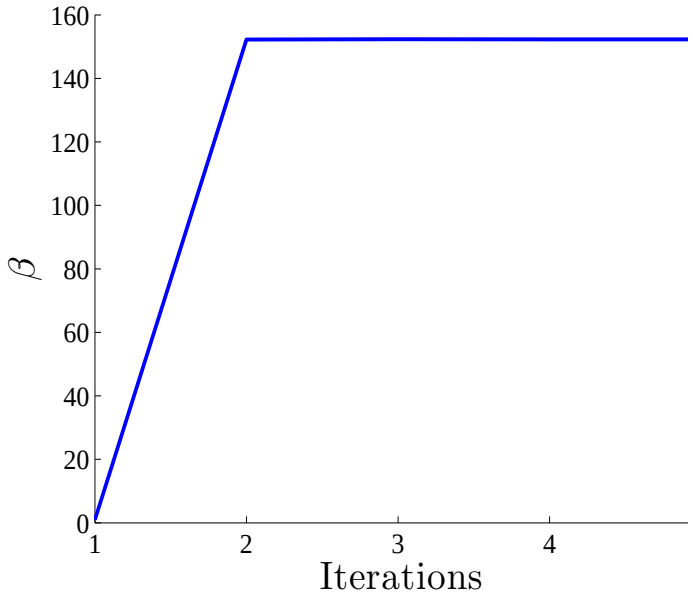


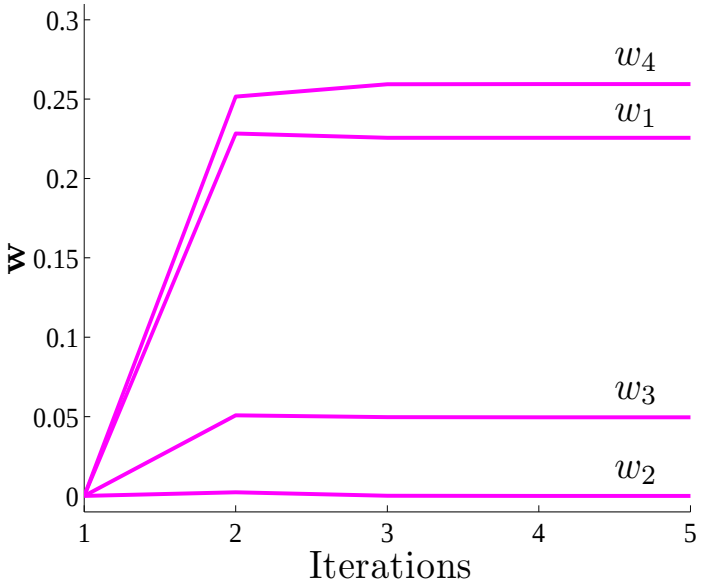
Итерационный процесс для матрицы Гессе (случай шумового параметра)

Случай шумовых признаков



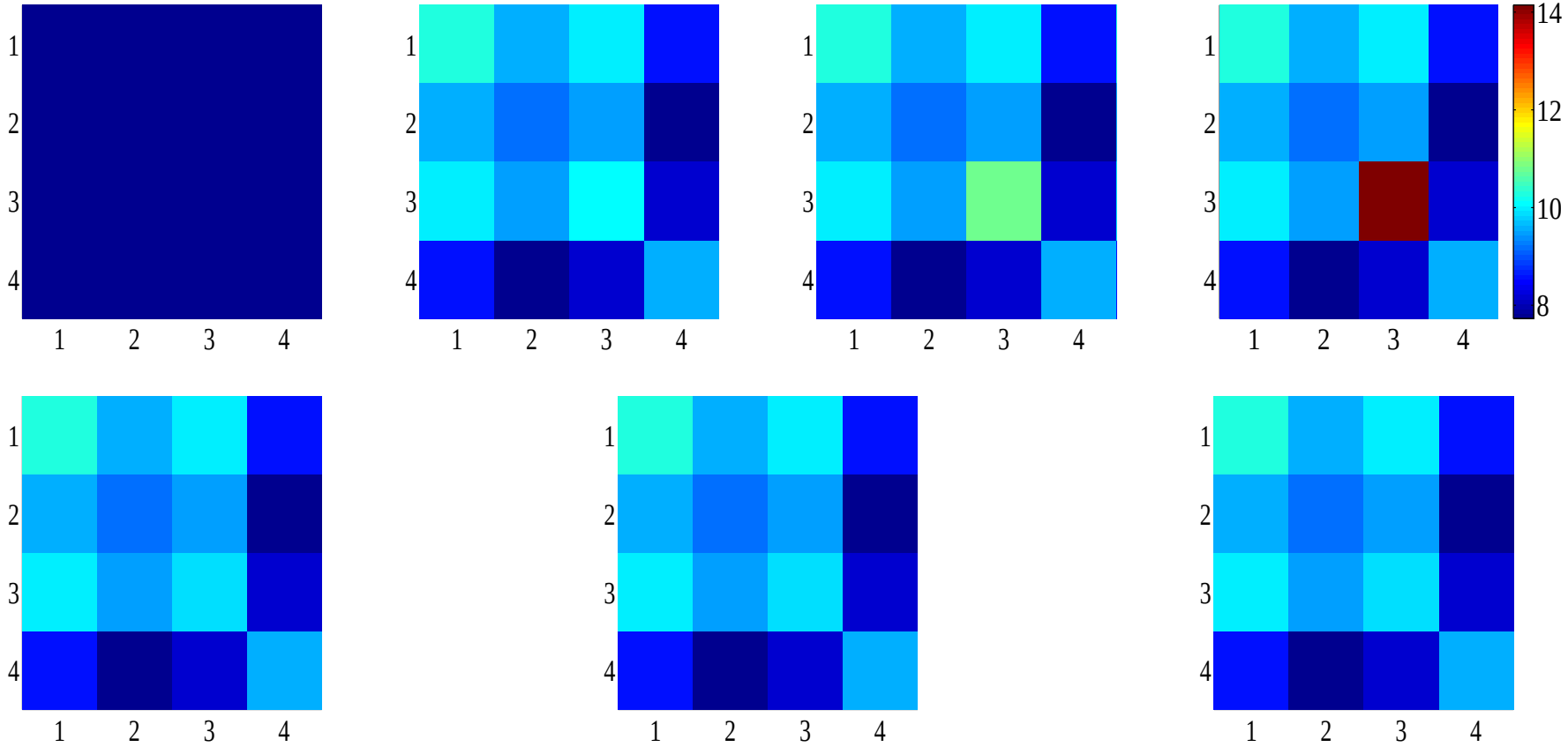






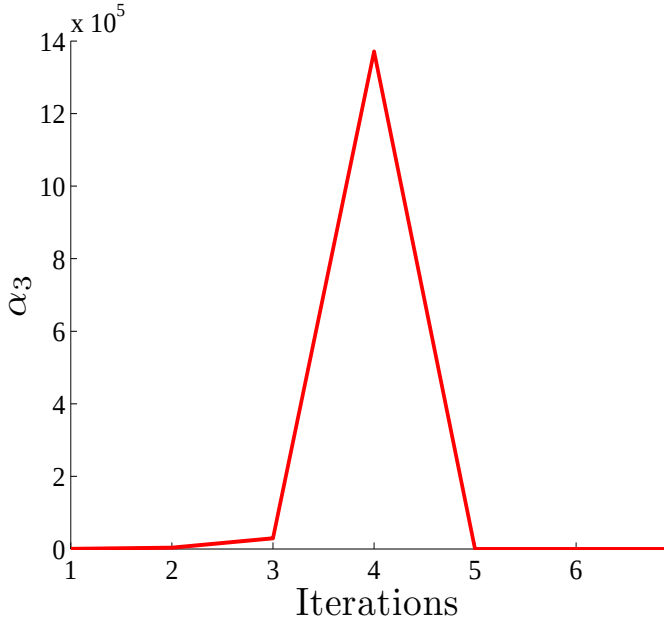
Случай коррелирующих признаков

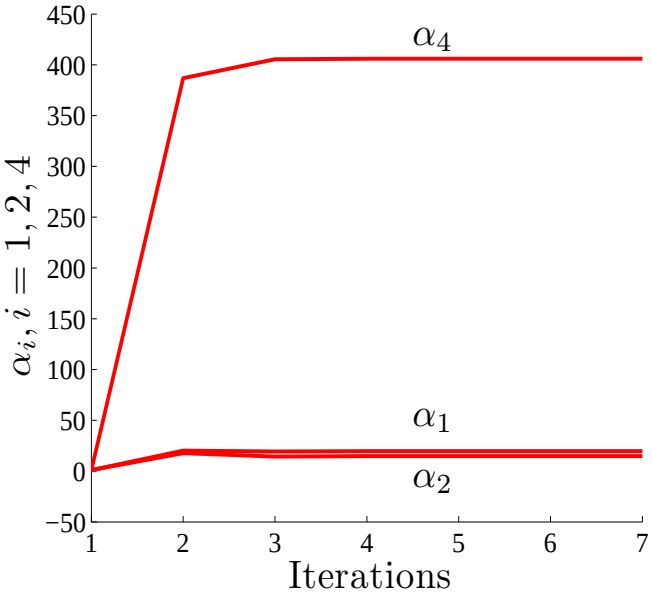
$\chi_1 = \xi^0$; $\chi_2 = \xi^1$; $\chi_3 = \xi^1 + k$, где $k \sim \mathcal{U}(0, 0.5)$; $\chi_4 = \sin(10\xi)$.

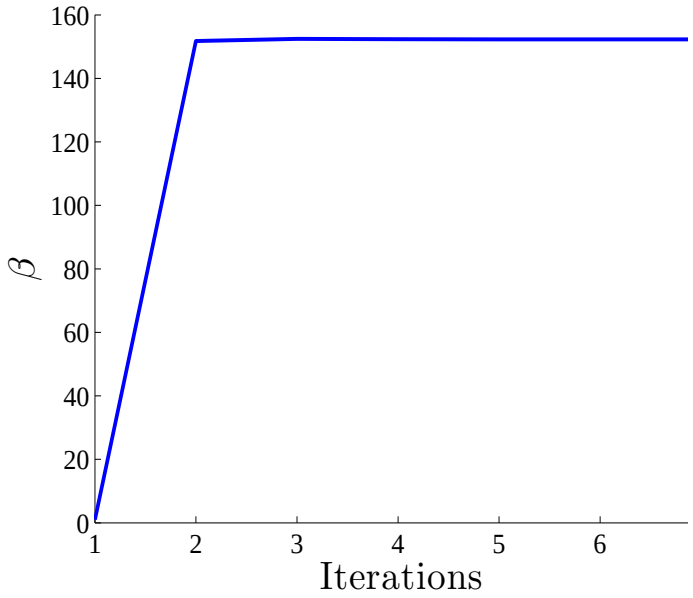


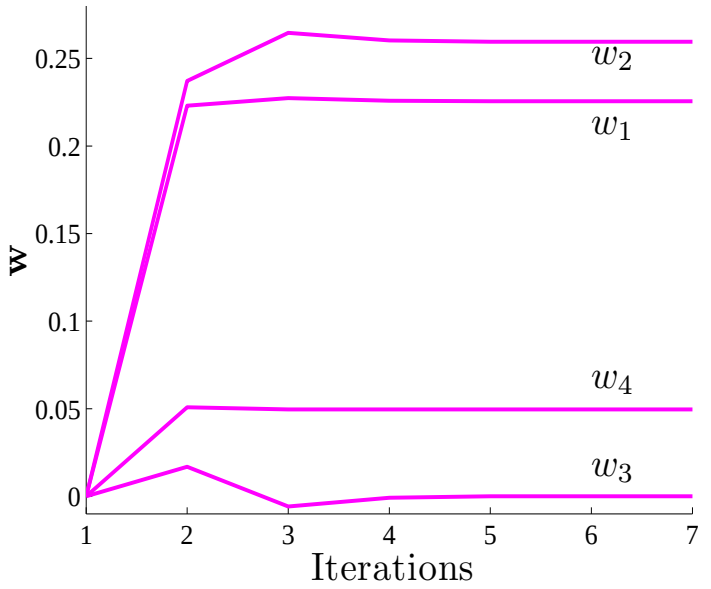
Итерационный процесс для матрицы Гессе (случай коррелирующих параметров модели)

Случай коррелирующих признаков









Результаты

1. предложен алгоритм, производящий фильтрацию шумовых и коррелирующих признаков;
2. модифицирован алгоритм Левенберга-Марквардта;
3. выполнен вычислительный эксперимент, иллюстрирующий работу алгоритмов.

Публикации

1. Стрижов В.В., Токмакова А.А. Оценивание гиперпараметров линейных регрессионных моделей при отборе шумовых и коррелирующих признаков // Информатика и её применения. — 2012. — № 4. — ISSN 1992-2264 (принято в печать).
2. Токмакова А.А. Получение устойчивых оценок гиперпараметров линейных регрессионных моделей // Машинное обучение и анализ данных. — 2011. — № 2. — С. 140-155. — ISSN 2223-3792.