ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2021, том 61, № 7, с. 1149–1161

ИНФОРМАТИКА

УДК 519.72

АНАЛИЗ ВЫБОРА АПРИОРНОГО РАСПРЕДЕЛЕНИЯ ДЛЯ СМЕСИ ЭКСПЕРТОВ[1)]

© 2021 г. А. В. Грабовой[1,*], В. В. Стрижов[1,2,**]

[1] 141701 Долгопрудный, М.о., Институтский пер., 9, Московский физико-технический институт, Россия

[2] 119333 Москва, ул. Вавилова, 40, ВЦ РАН им. А.А. Дородницына ФИЦ ИУ РАН, Россия

*e-mail: grabovoy.av@phystech.edu

**e-mail: strijov@phystech.edu

Поступила в редакцию 26.11.2020 г.
Переработанный вариант 26.11.2020 г.
Принята к публикации 11.03.2021 г.

Исследуются свойства смеси экспертов. Смесь экспертов – это ансамбль локальных аппроксимирующих моделей, которые являются экспертами и шлюзовой функцией, которая взвешивает данные экспертов. В качестве экспертов рассматриваются линейные модели, а в качестве шлюзовой функции – нейронная сеть с функцией softmax на последнем слое. Анализируются разные априорные распределения для каждого эксперта. Предложен метод, который учитывает связь между априорными распределениями разных экспертов. Для поиска оптимальных параметров локальных моделей и шлюзовой функции используется EM-алгоритм. Рассматривается задача распознавания окружностей на изображении. Каждый эксперт аппроксимирует одну окружность на изображении: находит координаты центра окружности и радиус окружности. Для анализа предложенного метода проводится вычислительный эксперимент на синтетических и реальных данных. В качестве реальных данных используются изображения радужки глаза, которые применяются в задачах распознавания радужки глаза. Библ. 23. Фиг. 13. Табл. 1.

Ключевые слова: смесь экспертов, байесовский выбор модели, априорное распределение.

DOI: 10.31857/S0044466921070073

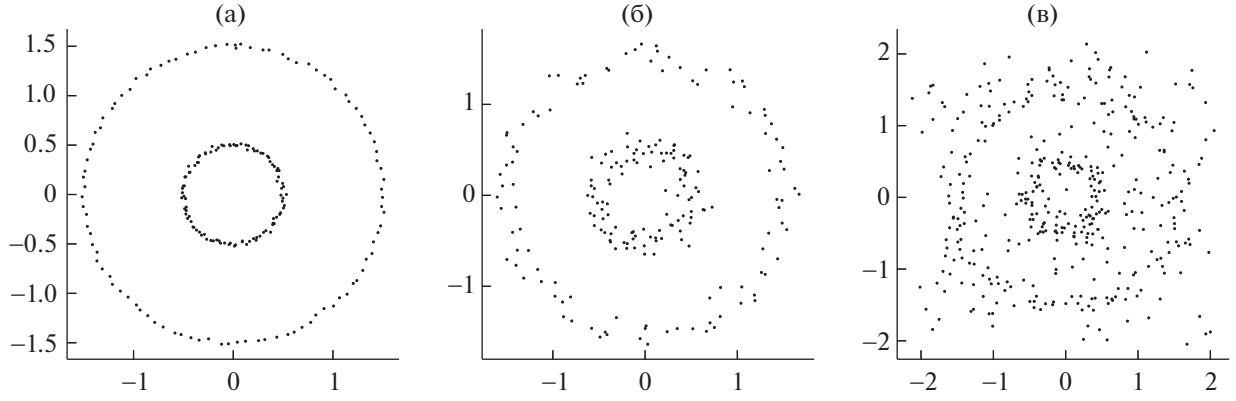
1. ВВЕДЕНИЕ

В работе исследуется проблема построения смеси экспертов. Смесь экспертов – это мультимодель, которая состоит из множества локальных моделей, называемых экспертами и шлюзовой функцией. Смесь экспертов использует шлюзовую функцию для взвешивания прогнозов каждого эксперта. Весовые коэффициенты шлюзовой функции зависят от объекта, для которого проводится прогноз. Примерами мультимоделей являются бэггинг, градиентный бустинг (см. [1]) и случайный лес (см. [2]). В [3] предполагается, что вклад каждого эксперта в ответ зависит от объекта из набора данных.

Основной проблемой построения мультимоделей является то, что ансамбль зависит от начальной инициализации параметров. Для улучшения устойчивости мультимодели предлагается использовать вероятностную постановку задачи для поиска оптимальных параметров шлюзовой функции и параметров локальной модели. В данной работе задается априорное распределение на параметры локальных моделей, а также предлагается учесть зависимость априорных распределений для разных моделей.

В настоящей работе решается задача поиска окружностей на бинаризованном изображении. Предполагается, что радиусы окружностей различаются значимо, а также, что центры почти сов-

[1)]Настоящая статья содержит результаты проекта Математические методы интеллектуального анализа больших данных, выполняемого в рамках реализации Программы Центра компетенций Национальной технологической инициативы “Центр хранения и анализа больших данных”, поддерживаемого Министерством науки и высшего образования по договору МГУ им. М.В. Ломоносова с Фондом поддержки проектов Национальной технологической инициативы от 11.12.2018 № 13/1251/2018. Работа выполнена при финансовой поддержке РФФИ (проекты 19-07-01155, 19-07-00875).



Фиг. 1. Пример окружностей с разным уровнем шума: (а) – окружности без шума, (б) – окружности с зашумленным радиусом, (в) – окружности с зашумленным радиусом, а также с равномерным шумом по всему изображению.

падают. Пример изображений показан на фиг. 1. В качестве экспертов рассматриваются линейные модели – каждая модель аппроксимирует одну окружность. В качестве шлюзовой функции рассматривается двухслойная нейронная сеть.

Большое количество работ в области построения смеси экспертов посвящены выбору шлюзовой функции: используется softmax, процесс Дирихле (см. [4]), нейронная сеть (см. [5]) с функцией softmax на последнем слое. Ряд работ посвящен выбору моделей в качестве отдельных экспертов. В качестве модели эксперта в [6], [7] рассматривается линейная модель, в [8], [9] – модель SVM. В [3] представлен обзор методов и моделей в задачах смеси экспертов.

Смесь экспертов имеет множество приложений в прикладных задачах. Работы [10]–[12] посвящены применению смеси экспертов в задачах прогнозирования временных рядов. В [13] предложен метод распознавания рукописных цифр. Метод распознавания текстов с помощью смеси экспертов иследуется в [14], распознавание речи – в [15]–[17]. В [18] исследуется смесь экспертов для задачи распознавания трехмерных движений человека. В [19] описаны работы по исследованию обнаружения радужки глаза на изображении. В [20], [21], в частности, описаны методы выделения границ радужки и зрачка.

2. ПОСТАНОВКА ЗАДАЧИ АППРОКСИМАЦИИ ПАРАМЕТРОВ ОКРУЖНОСТИ

Задано бинарное изображение

$$\mathbf{M} \in \{0, 1\}^{m_1 \times m_2},$$

где 1 – это черный пиксель, который принадлежит рассматриваемой фигуре на изображении, а 0 – белый пиксель, который является фоном изображения. Пример изображения показан на фиг. 1.

Изображение $\mathbf{M}$ отображается в множество координат $\mathbf{C} = \{x_i, y_i\}_{i=1}^{N}$. Координата (x_i, y_i) является координатой i-го черного пикселя на изображении $\mathbf{M}$:

$$\mathbf{C} \in \mathbb{R}^{N \times 2},$$

где N – число черных пикселей.

Обозначим через (x_0, y_0) центр окружности, а r – радиус окружности. Координаты $(x_i, y_i) \in \mathbf{C}$ – это геометрическое место точек, которое удовлетворяет системе уравнений

$$(x_i - x_0)^2 + (y_i - y_0)^2 = r^2 + \varepsilon_i, \quad i \in \{1, 2, \ldots, N\},$$

где $\varepsilon_i \in \mathcal{N}(0, \beta^{-1})$ – невязка i-го уравнения, которая является следствием шума на изображении. Раскрыв скобки, получим

$$(2x_0) \cdot x_i + (2y_0) \cdot y_i + (r^2 - x_0^2 - y_0^2) \cdot 1 = x_i^2 + y_i^2 - \varepsilon_i. \tag{2.1}$$

Выражение (2.1) переписывается в задачу линейной регрессии следующим образом:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbf{R}^n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2, \quad \mathbf{X} = [\mathbf{C}, 1], \quad \mathbf{y} = [x_1^2 + y_1^2, \ldots, x_N^2 + y_N^2]^{\mathrm{T}}. \tag{2.2}$$

Используя вектор параметров $\mathbf{w} = [w_1, w_2, w_3]^{\mathrm{T}}$, получаем параметры окружности x_0, y_0, r:

$$x_0 = \frac{w_1}{2}, \quad y_0 = \frac{w_2}{2}, \quad r = \sqrt{w_3 + x_0^2 + y_0^2}.$$

Решая уравнения (2.2), находим параметры единственной окружности на изображении. В случае, когда на изображении несколько окружностей, предлагается использовать смесь экспертов, которая состоит из линейных моделей – экспертов. Каждый эксперт описывает одну окружность на изображении.

3. ПОСТАНОВКА ЗАДАЧИ ПОСТРОЕНИЯ СМЕСИ ЭКСПЕРТОВ

Обобщим подход аппроксимации одной окружности на изображении на случай, когда на изображении несколько окружностей. Пусть изображение состоит из K окружностей, тогда множество черных пикселей $\mathbf{C}$ представляется в виде

$$\mathbf{C} = \coprod_{k=1}^{K} \mathbf{C}'_k,$$

где $\mathbf{C}'_k$ – множество точек, принадлежащих k-й окружности. Множеству точек $\mathbf{C}'_k \subset \mathbf{C}$ соответсвует задача линейной регрессии для выборки $\mathbf{X}'_k \subset \mathbf{X}$, $\mathbf{y}'_k \subset \mathbf{y}$. Модель $\mathbf{g}_k$, аппроксимирующая k-ю подвыборку $\mathbf{X}'_k$, $\mathbf{y}'_k$, является локальной моделью для выборки $\mathbf{X}$, $\mathbf{y}$.

Определение 1. Модель $\mathbf{g}$ называется *локальной моделью* для выборки $\mathbf{X}$, $\mathbf{y}$, если $\mathbf{g}$ аппроксимирует некоторое непустое подмножество $\mathbf{X}'$, $\mathbf{y}'$ этой выборки.

Определение 2. Мультимодель $\mathbf{f}$ называется *смесью экспертов*, если

$$\mathbf{f} = \sum_{k=1}^{K} \pi_k \mathbf{g}_k(\mathbf{w}_k), \quad \pi_k(\mathbf{x}, \mathbf{V}) : \mathbb{R}^{n\times|\mathbf{V}|} \to [0, 1], \quad \sum_{k=1}^{K} \pi_k(\mathbf{x}, \mathbf{V}) = 1, \tag{3.1}$$

где $\mathbf{g}_k$ является k-й локальной моделью, π_k – шлюзовая функция, вектор $\mathbf{w}_k$ – параметр k-й локальной модели, а $\mathbf{V}$ – параметры шлюзовой функции.

В данной работе в качестве локальных моделей рассматриваются линейные модели. В качестве шлюзовой функции рассматривается двухслойный перцептрон:

$$\mathbf{g}_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x}, \quad \pi(\mathbf{x}, \mathbf{V}) = \mathbf{softmax}(\mathbf{V}_1^{\mathrm{T}}\sigma(\mathbf{V}_2^{\mathrm{T}}\mathbf{x})), \tag{3.2}$$

где $\mathbf{V} = \{\mathbf{V}_1, \mathbf{V}_2\}$ – множество параметров шлюзовой функции.

Предлагается использовать вероятностный подход для описания смеси экспертов. Вводится предположение, что $\mathbf{y}$ является случайным вектором, который задается плотностью распределения $p(\mathbf{y}|\mathbf{X})$. Предполагается, что плотность распределения $p(\mathbf{y}|\mathbf{X}, \mathbf{f})$ аппроксимирует истинную плотность распределения $p(\mathbf{y}|\mathbf{X})$:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{f}) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k p_k(y_i|\mathbf{g}_k(\mathbf{x}_i)) \right), \tag{3.3}$$

где $\mathbf{f}$ – смесь экспертов, а $\mathbf{g}_k, \pi$ определяются выражением (3.2).

Пусть $\mathbf{w}_k$ является случайным вектором, который задается плотностью распределения $p^k(\mathbf{w}_k)$. Получим совместное распределение параметров локальных моделей и вектора ответов:

$$p(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}) = \prod_{k=1}^{K} p^k(\mathbf{w}_k) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k p_k(y_i|\mathbf{w}_k, \mathbf{x}_i) \right), \tag{3.4}$$

где $\mathbf{W} = \{\mathbf{w}_1, \ldots, \mathbf{w}_K\}$. Оптимальные параметры находятся с помощью максимизации правдоподобия:

$$\hat{\mathbf{V}}, \hat{\mathbf{W}} = \arg\max_{\mathbf{V},\mathbf{W}} p(\mathbf{y}, \mathbf{W} | \mathbf{X}, \mathbf{V}).$$

4. ВЕРОЯТНОСТНАЯ ПОСТАНОВКА СМЕСИ ЭКСПЕРТОВ

Для построения смеси экспертов (3.1), (3.4) введем следующие вероятностные предположения о данных (2.2):

(i) правдоподобие $p_k(y_i | \mathbf{w}_k, \mathbf{x}_i) = \mathcal{N}(y_i | \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i, \beta^{-1})$, где параметр β является уровнем шума,

(ii) априорное распределение параметров $p^k(\mathbf{w}_k) = \mathcal{N}(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k)$, где $\mathbf{w}_k^0$ – вектор размерности $n \times 1$, а $\mathbf{A}_k$ – ковариационная матрица размерности $n \times n$,

(iii) регуляризация априорного распределения $p(\varepsilon_{k,k'} | \Xi) = \mathcal{N}(\varepsilon_{k,k'} | 0, \Xi)$, где Ξ – ковариационная матрица, а $\varepsilon_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0$.

Предположение (i) задает априорное предположение о распределениии вектора параметров локальной модели $\mathbf{w}_k$. Априорное распределение задает ограничения на локальную модель. Например, если $\mathbf{w}_k^0 = [0, 0, 1]$, то k-я локальная модель аппроксимирует окружность с параметрами $x_0 = 0$, $y_0 = 0$, $r = 1$ с большей вероятностью.

Предположение (iii) задает регуляризацию априорных распределений. Она учитывает связь между априорными ограничениями разных локальных моделей. Например, если $\mathrm{diag}(\Xi) = [0.001, 0.001, 1]$, то центры разных окружностей совпадают.

Используя предположения (i)–(iii) и выражение (3.4), получаем полное правдоподобие:

$$p(\mathbf{y}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k N(y_i | \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i, \beta^{-1}) \right) \prod_{k=1}^{K} \mathcal{N}(\mathbf{w}_k | \mathbf{w}_k^0, \mathbf{A}_k) \prod_{k,k'=1}^{K} \mathcal{N}(\varepsilon_{k,k'} | 0, \Xi), \tag{4.1}$$

где $\mathbf{A} = \{\mathbf{A}_1, \cdots, \mathbf{A}_K\}$.

Введем бинарную матрицу $\mathbf{Z}$. Элемент матрицы $z_{ik} = 1$ тогда и только тогда, когда i-й объект аппроксимируется k-й локальной моделью. Подставляя бинарную матрицу $\mathbf{Z}$ в выражении (4.1), а также взяв логарифм, получаем

$$\begin{aligned} \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) = {} & \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} (y_i - \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\ & + \sum_{k=1}^{K} \left[-\frac{1}{2} (\mathbf{w}_k - \mathbf{w}_k^0)^{\mathrm{T}} \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] + \\ & + \sum_{k=1}^{K} \sum_{k'=1}^{K} \left[-\frac{1}{2} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0)^{\mathrm{T}} \Xi^{-1} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0) + \frac{1}{2} \log \det \Xi - \frac{n}{2} \log 2\pi \right]. \end{aligned} \tag{4.2}$$

Получаем новую задачу оптимизации обоснованности. Функция обоснованности получается при интегрировании выражения (4.2) по параметрам $\mathbf{W}$, $\mathbf{Z}$:

$$\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta = \arg\max_{\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta} \int_{\mathbf{W},\mathbf{Z}} \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) d\mathbf{W} d\mathbf{Z}. \tag{4.3}$$

5. ЕМ-АЛГОРИТМ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ОПИМИЗАЦИИ

Рассмотрим вариационную плотность $q(\mathbf{W}, \mathbf{Z})$ для параметров $\mathbf{W}$, $\mathbf{Z}$. Тогда функция обоснованности принимает следующий вид:

$$\begin{aligned} \log p(\mathbf{y} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) &= \int_{\mathbf{W},\mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \log p(\mathbf{y} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) d\mathbf{W} d\mathbf{Z} = \\ &= \int_{\mathbf{W},\mathbf{Z}} q(\mathbf{W}, \mathbf{Z}) \log \frac{p(\mathbf{y}, \mathbf{W}, \mathbf{Z} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta)}{p(\mathbf{W}, \mathbf{Z} | \mathbf{y}, \mathbf{X}, \mathrm{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta)} d\mathbf{W} d\mathbf{Z} = \end{aligned}$$

$$
\begin{aligned}
&= \int\limits_{\mathbf{W},\mathbf{Z}} q(\mathbf{W},\mathbf{Z}) \log \frac{p(\mathbf{y},\mathbf{W},\mathbf{Z}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\Xi,\beta) q(\mathbf{W},\mathbf{Z})}{p(\mathbf{W},Z|\mathbf{y},\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\Xi,\beta) q(\mathbf{W},\mathbf{Z})} d\mathbf{W} d\mathbf{Z} = \\
&= \int\limits_{\mathbf{W},\mathbf{Z}} q(\mathbf{W},\mathbf{Z}) \frac{p(\mathbf{y},\mathbf{W},\mathbf{Z}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\Xi,\beta)}{q(\mathbf{W},\mathbf{Z})} d\mathbf{W} d\mathbf{Z} + \\
&\quad + \int\limits_{\mathbf{W},\mathbf{Z}} q(\mathbf{W},\mathbf{Z}) \frac{q(\mathbf{W},\mathbf{Z})}{p(\mathbf{W},\mathbf{Z}|\mathbf{y},\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\Xi,\beta)} d\mathbf{W} d\mathbf{Z} = \\
&= \mathcal{L}(q,\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta) + D_{KL}(q(\mathbf{W},\mathbf{Z}) \| p(\mathbf{W},\mathbf{Z}|\mathbf{y},\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\Xi,\beta)).
\end{aligned} \tag{5.1}
$$

Используя (5.1), получаем нижнюю оценку обоснованности:

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\Xi,\beta) \geq \mathcal{L}(q,\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta),$$

где $\mathcal{L}(q,\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta)$ называется нижней оценкой обоснованности.

Используем EM-алгоритм (см. [22], [23]) для решения оптимизационной задачи (4.3). Заметим, что EM-алгоритм вместо оптимизации $\log p(\mathbf{y}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\Xi,\beta)$ оптимизирует нижнюю оценку $\mathcal{L}(q,\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta)$.

E-шаг. E-шаг решает следующую оптимизационую задачу:

$$\mathcal{L}(q,\mathbf{V},\mathbf{W}^0,\mathbf{A},\beta) \to \max_{q(\mathbf{W},\mathbf{Z})},$$

где параметры $\mathbf{V}$, $\mathbf{W}^0$, $\mathbf{A}$, β являются зафиксированными.

Пусть совместное распределение $q(\mathbf{Z},\mathbf{W})$ удовлетворяет условию независимости $q(\mathbf{Z},\mathbf{W}) = q(\mathbf{Z})q(\mathbf{W})$ (см. [23]). Далее символом $\propto$ обозначим то, что обе стороны выражения равны с точностью до аддитивной константы. Сначала найдем распределение $q(\mathbf{Z})$:

$$
\begin{aligned}
&\log q(\mathbf{Z}) = \mathrm{E}_{q/\mathbf{Z}} \log p(\mathbf{y},\mathbf{Z},\mathbf{W}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\Xi,\beta) \propto \\
&\propto \sum_{i+1}^{N}\sum_{k=1}^{K} z_{ik}\left[\log \pi_k(\mathbf{x}_i,\mathbf{V}) - \frac{\beta}{2}(y_i^2 - \mathbf{x}_i^{\mathrm{T}}\mathrm{E}\mathbf{w}_k + \mathbf{x}_i^{\mathrm{T}}\mathrm{E}\mathbf{w}_k\mathbf{w}_k^{\mathrm{T}}\mathbf{x}_i) + \frac{1}{2}\log\frac{\beta}{2\pi}\right], \\
&p(z_{ik}=1) = \frac{\exp\left[\log \pi_k(\mathbf{x}_i,\mathbf{V}) - \frac{\beta}{2}(\mathbf{x}_i^{\mathrm{T}}\mathrm{E}\mathbf{w}_k\mathbf{w}_k^{\mathrm{T}}\mathbf{x}_i - \mathbf{x}_i^{\mathrm{T}}\mathrm{E}\mathbf{w}_k)\right]}{\sum\limits_{k'=1}^{K}\exp\left[\log \pi_{k'}(\mathbf{x}_i,\mathbf{V}) - \frac{\beta}{2}(\mathbf{x}_i^{\mathrm{T}}\mathrm{E}\mathbf{w}_{k'}\mathbf{w}_{k'}^{\mathrm{T}}\mathbf{x}_i - \mathbf{x}_i^{\mathrm{T}}E\mathbf{w}_{k'})\right]}.
\end{aligned} \tag{5.2}
$$

Используя выражения (5.2), получаем, что распределение $q(z_{ik})$ является бернулевским распределением с параметром z_{ik}, которое задается выражением (5.2). Далее найдем распределение $q(\mathbf{W})$:

$$
\begin{aligned}
&\log q(\mathbf{W}) = \mathrm{E}_{q/\mathbf{W}} \log p(\mathbf{y},\mathbf{Z},\mathbf{W}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\Xi,\beta) \propto \\
&\propto \sum_{i=1}^{N}\sum_{k=1}^{K} \mathrm{E}z_{ik}\left[\log \pi_k(\mathbf{x}_{i,\mathbf{V}}) - \frac{\beta}{2}(y_i - \mathbf{w}_k^{\mathrm{T}}\mathbf{x}_i)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\
&+ \sum_{k=1}^{K}\left[-\frac{1}{2}(\mathbf{w}_k - \mathbf{w}_k^0)^{\mathrm{T}}\mathbf{A}_k^{-1}(\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] \propto \\
&\propto \sum_{k=1}^{K}\left[\mathbf{w}_k^{\mathrm{T}}\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0 + \beta\sum_{i=1}^{N}\mathbf{x}_i y_i \mathrm{E}z_{ik}\right) - \frac{1}{2}\mathbf{w}_k^{\mathrm{T}}\left(\mathbf{A}_k^{-1} + \beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{w}_k\right].
\end{aligned} \tag{5.3}
$$

Используя выражение (5.3), получаем, что распределение $q(\mathbf{w}_k)$ является нормальным распределением со средним $\mathbf{m}_k$ и ковариационной матрицей $\mathbf{B}_k$:

$$\mathbf{m}_k = \mathbf{B}_k\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0 + \beta\sum_{i=1}^{N}\mathbf{x}_i y_i \mathrm{E}z_{ik}\right), \quad \mathbf{B}_k = \left(\mathbf{A}_k^{-1} + \beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathrm{E}z_{ik}\right)^{-1}.$$

М-шаг. М-шаг решает следующую оптимизационную задачу:

$$\mathscr{L}(q, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta) \to \max_{\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta},$$

где $q(\mathbf{W}, \mathbf{Z})$ является известной плотностью распределения. Распределение $q(\mathbf{Z}, \mathbf{W})$ является фиксированным, в то время как вариационная нижняя оценка $\mathscr{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)$ максимизируется по параметрам $\mathbf{V}$, $\mathbf{W}^0$, $\mathbf{A}$, β:

$$\begin{aligned}
&\mathscr{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta) = \mathrm{E}_q \log p(\mathbf{y}, \mathbf{Z}, \mathbf{W} \,|\, \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \Xi, \beta) = \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} \mathrm{E} z_{ik} \left[\log \pi_k(\mathbf{x}_i, \mathbf{V}) - \frac{\beta}{2} \mathrm{E}(y_i - \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i)^2 + \frac{1}{2} \log \frac{\beta}{2\pi} \right] + \\
&+ \sum_{k=1}^{K} \left[-\frac{1}{2} \mathrm{E}(\mathbf{w}_k - \mathbf{w}_k^0)^{\mathrm{T}} \mathbf{A}_k^{-1} (\mathbf{w}_k - \mathbf{w}_k^0) + \frac{1}{2} \log \det \mathbf{A}_k^{-1} - \frac{n}{2} \log 2\pi \right] + \\
&+ \sum_{k=1}^{K} \sum_{k'=1}^{K} \left[-\frac{1}{2} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0)^{\mathrm{T}} \Xi^{-1} (\mathbf{w}_k^0 - \mathbf{w}_{k'}^0) + \frac{1}{2} \log \det \Xi - \frac{n}{2} \log 2\pi \right].
\end{aligned} \tag{5.4}$$

Для нахождения оптимального параметра $\mathbf{V}$ используется градиентный метод оптимизации, который сходится к некоторому локальному экстремуму. Используя выражения (5.4), получаем оптимальное значение параметра $\mathbf{A}_k$:

$$\begin{aligned}
&\frac{\partial \mathscr{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \mathbf{A}_k^{-1}} = \frac{1}{2} \mathbf{A}_k - \frac{1}{2} \mathrm{E}(\mathbf{w}_k - \mathbf{w}_k^0)(\mathbf{w}_k - \mathbf{w}_k^0)^{\mathrm{T}} = 0, \\
&\mathbf{A}_k = \mathrm{E}\mathbf{w}_k \mathbf{w}_k^{\mathrm{T}} - \mathbf{w}_k^0 \mathrm{E}\mathbf{w}_k^{\mathrm{T}} - \mathrm{E}\mathbf{w}_k \mathbf{w}_k^{0\mathrm{T}} + \mathbf{w}_k^0 \mathbf{w}_k^{0\mathrm{T}}.
\end{aligned}$$

Аналогично получаем оптимальные значения для параметра β и для параметров $\mathbf{w}_k^0$:

$$\begin{aligned}
&\frac{\partial \mathscr{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \beta} = \sum_{k=1}^{K} \sum_{i=1}^{N} \left(\frac{1}{\beta} \mathrm{E} z_{ik} - \frac{1}{2} \mathrm{E} z_{ik} [y_i^2 - 2 y_i \mathbf{x}_i^{\mathrm{T}} \mathrm{E}\mathbf{w}_k + \mathbf{x}_i^{\mathrm{T}} \mathbf{w}_k \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i] \right) = 0, \\
&\frac{1}{\beta} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} [y_i^2 - 2 y_i \mathbf{x}_i^{\mathrm{T}} \mathrm{E}\mathbf{w}_k + \mathbf{x}_i^{\mathrm{T}} \mathrm{E}\mathbf{w}_k \mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i] \mathrm{E} z_{ik}, \\
&\frac{\partial \mathscr{L}(\mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta)}{\partial \mathbf{w}_k^0} = \mathbf{A}_k^{-1} (\mathrm{E}\mathbf{w}_k - \mathbf{w}_k^0) + \Xi \sum_{k'=1}^{K} [\mathbf{w}_{k'}^0 - \mathbf{w}_k^0] = 0, \\
&\mathbf{w}_k^0 = [\mathbf{A}_k^{-1} + (K-1)\Xi]^{-1} \left(\mathbf{A}_k^{-1} \mathrm{E}\mathbf{w}_k + \Xi \sum_{k'=1, k' \neq k}^{K} \mathbf{w}_{k'}^0 \right).
\end{aligned} \tag{5.5}$$

Выражения (5.2)–(5.5) задают итеративную процедуру, которая сходится к некоторому локальному максимуму оптимизационной задачи (4.3).

6. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Для анализа качества различных мультимоделей для аппроксимации окружности проводится вычислительный эксперимент. В эксперименте рассматриваются следующие мультимодели: мультимодель $\mathbf{f}_1$ без использования априорных распределений, мультимодель $\mathbf{f}_2$, которая использует априорные распределения (6.2) для параметров, и мультимодель $\mathbf{f}_3$, которая использует регуляризацию априорных распределений. Точность аппроксимации мультимодели $\mathbf{f}_i$ задается следующим образом:

$$\mathscr{S}_{\mathbf{f}_i} = \sum_{k=1}^{K} (x_0^k - x_{\mathrm{pr}}^k)^2 + (y_0^k - y_{\mathrm{pr}}^k)^2 + (r^k - r_{\mathrm{pr}}^k)^2, \tag{6.1}$$

где x_0^k, y_0^k, r^k – истинный центр и радиус для k-й окружности соответственно, x_{pr}^k, y_{pr}^k, r_{pr}^k – предсказанные центр и радиус для k-й окружности соответственно.

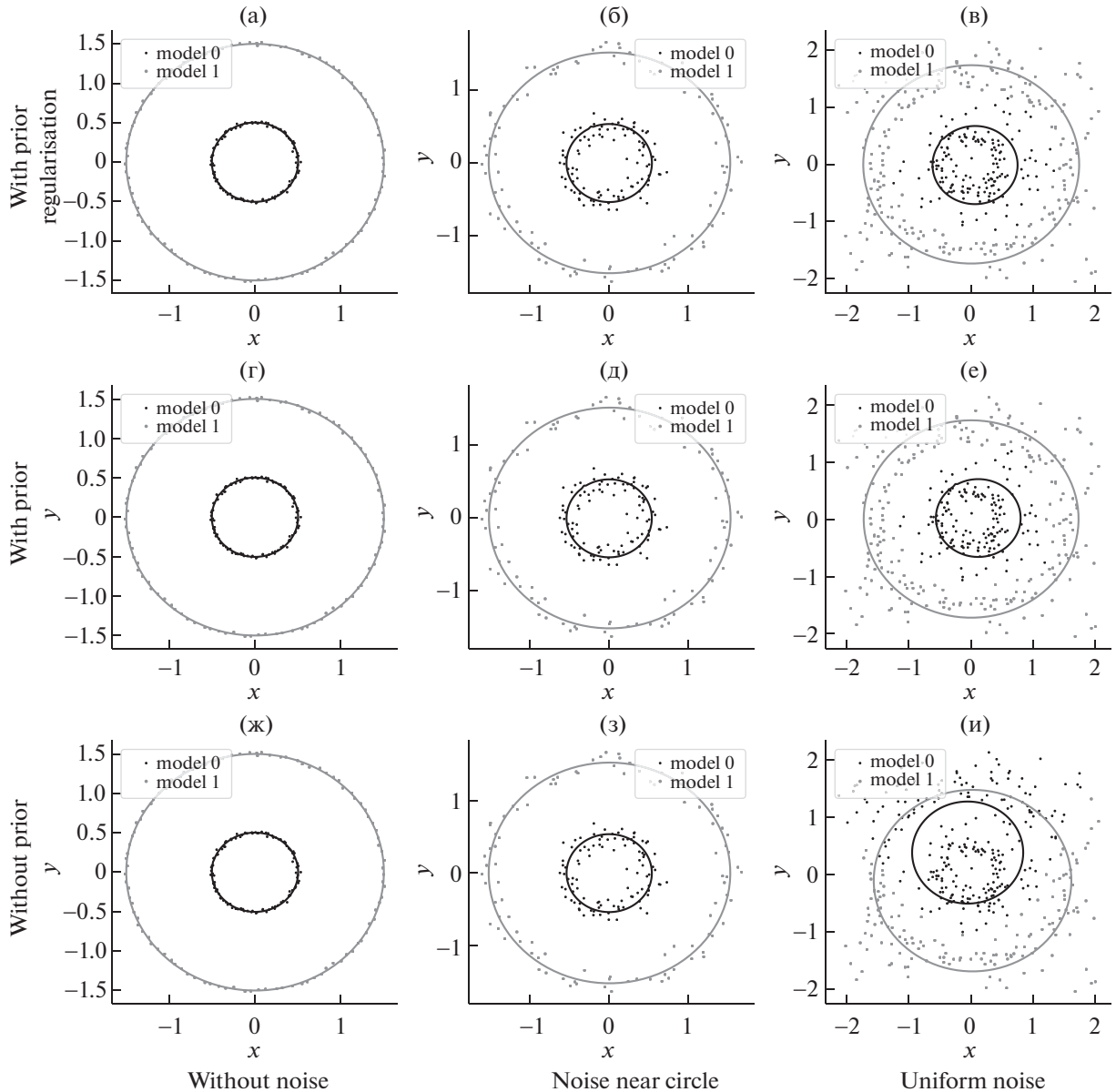
Таблица 1. Качество аппроксимации (6.1) для всех мультимоделей

Выборка	$\mathscr{S}_{\mathbf{f}_1}$	$\mathscr{S}_{\mathbf{f}_2}$	$\mathscr{S}_{\mathbf{f}_3}$
Synthetic 1	10^{-5}	10^{-5}	10^{-5}
Synthetic 2	0.6	10^{-3}	10^{-3}
Synthetic 3	0.6	10^{-3}	10^{-3}

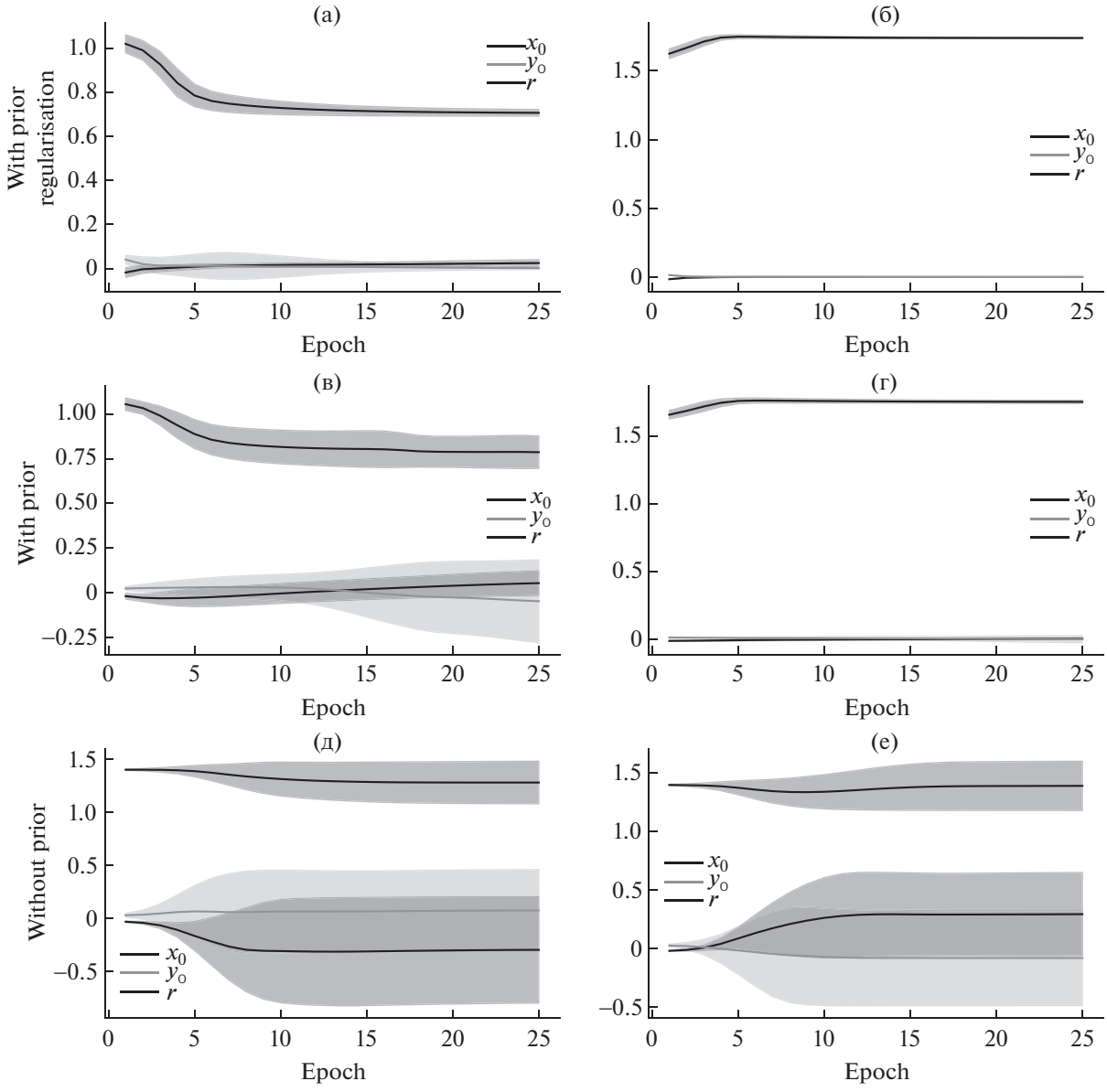
Для сравнения моделей с разными вероятностными предположениями используется правдоподобие (3.3). В вычислительном эксперименте используется следующее априорное распределение:

$$p^1(\mathbf{w}_1) \sim \mathcal{N}(\mathbf{w}_1^0, \mathbf{I}), \quad p^2(\mathbf{w}_2) \sim \mathcal{N}(\mathbf{w}_2^0, \mathbf{I}), \tag{6.2}$$

где $\mathbf{w}_1^0 = [0, 0, 0.1]$, $\mathbf{w}_2^0 = [0, 0, 2]$.



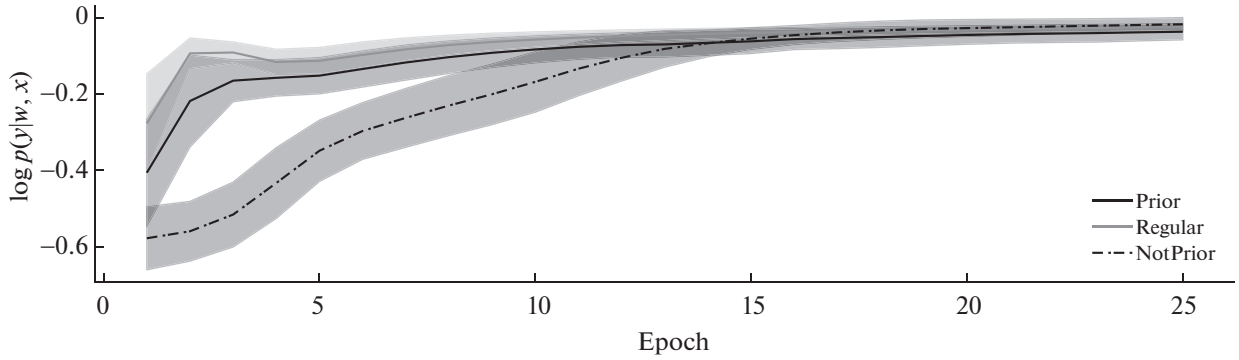
Фиг. 2. Мультимодель в зависимости от разных априорных предположений и в зависимости от разного уровня шума: (а)–(в) – модель с регуляризацией априорных распределений, (г)–(е) – модель с заданными априорными распределениями на параметрах локальных моделей, (ж)–(и) – модель без заданных априорных предположений.



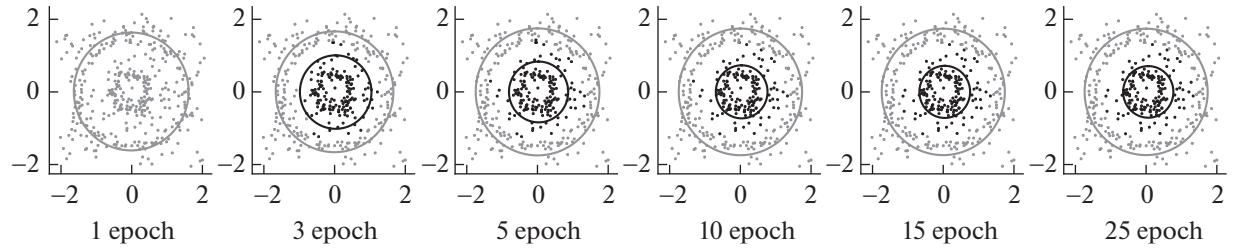
Фиг. 3. Зависимости центра и радиуса окружностей от номера итерации: (а), (б) – модель с регуляризацией априорных распределений; (в), (г) – модель с заданными априорными распределениями на параметры моделей; (д), (е) – модель без задания априорных распределений.

Синтетические данные с разным типом шума в изображении. В вычислительном эксперименте сравнивается качество следующих мультимоделей $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$ на синтетических данных. Синтетические данные являются двумя концентрическими окружностями с разным уровнем шума. Выборка Synthetic 1 является изображением без шума, выборка Synthetic 2 – изображением с зашумленным радиусом окружности, а выборка Synthetic 3 – изображением с равномерным шумом. На фиг. 2 показаны результаты для мельтимоделей $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$. Все модели оптимизировались с помощью 50 итераций EM-алгоритма. Мультимодели $\mathbf{f}_2$, $\mathbf{f}_3$ аппроксимируют окружности лучше, чем мультимодель $\mathbf{f}_1$. В табл. 1 показано качество аппрроксимации (6.1) для всех мультимоделей.

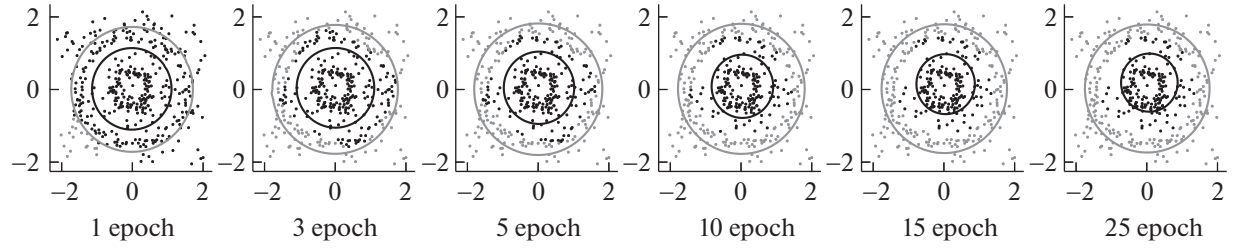
Анализ сходимости на синтетической выборке. Данная часть эксперимента анализирует качество сходимости EM-алгоритма для разных мультимоделей $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$. Анализ всех мультимоделей проводится на выборке Synthetic 3.



Фиг. 4. Зависимости логарифма правдоподобия (3.3) от номера итерации.



Фиг. 5. Визуализации процесса сходимости мультимодели с использованием априорной регуляризации.



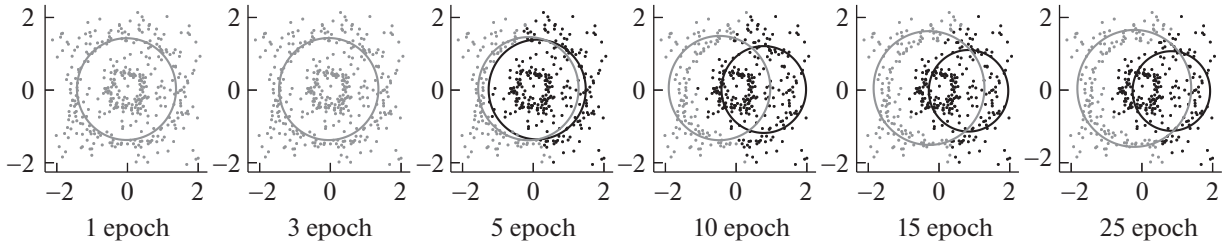
Фиг. 6. Визуализации процесса сходимости мультимодели с использованием априорного распределения.

На фиг. 3 показана зависимость предсказанных центра и радуса в зависимости от номера итерации EM-алгоритма. Мультимодель $\mathbf{f}_2$, которая использует априорное распределение, аппроксимирует окружность лучше мультимодели $\mathbf{f}_1$, которая не использует никакого априорного распределения. Мультимодель $\mathbf{f}_3$, которая использует регуляризатор априорных распределений, является более стабильной, чем мультимодель $\mathbf{f}_2$.

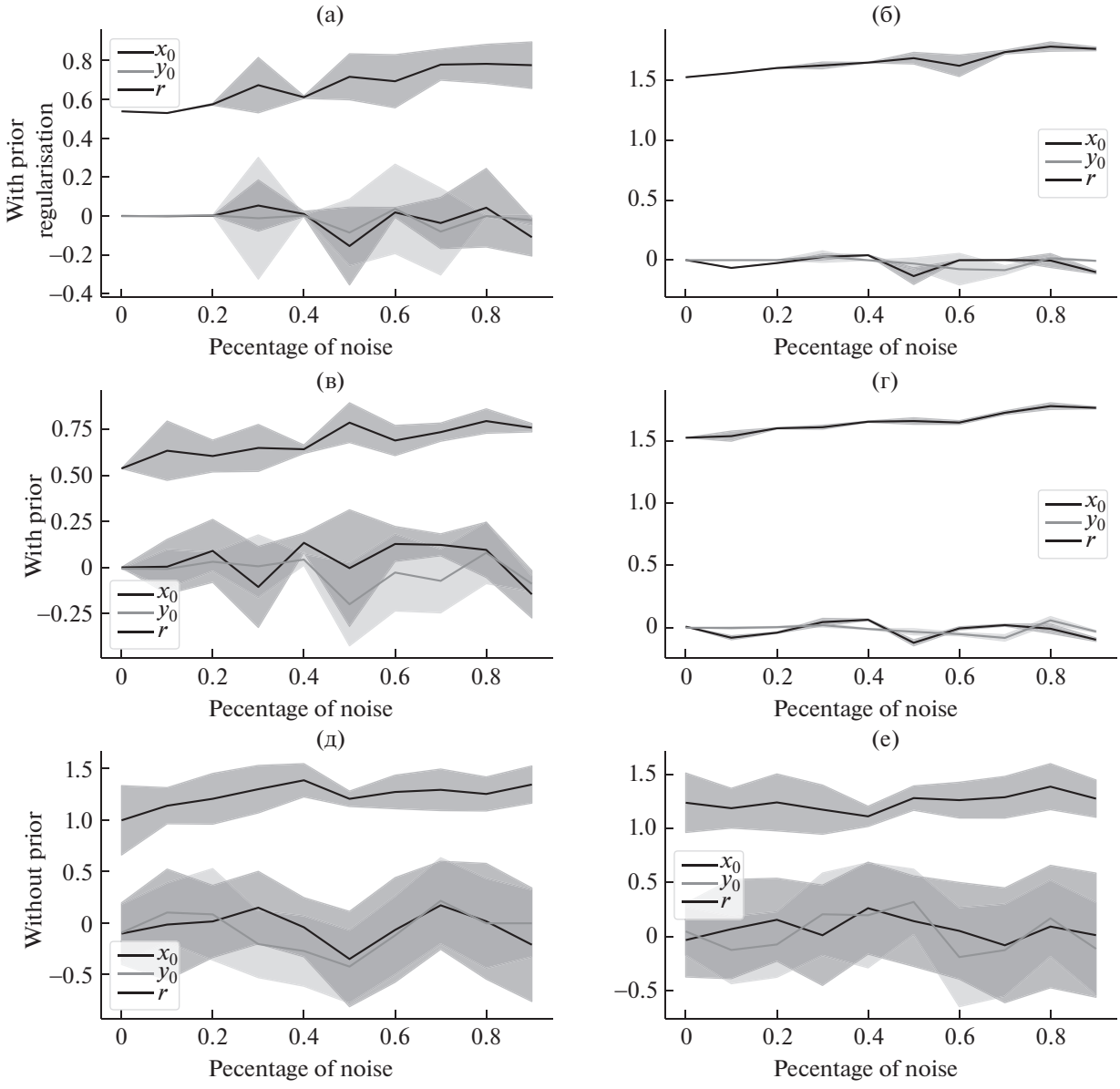
На фиг. 4 показана зависимость логарифма правдоподобия (3.3) от номера итерации EM-алгоритма. Логарифм правдоподобия мультимоделей $\mathbf{f}_2$, $\mathbf{f}_3\mathbf{f}_2, \mathbf{f}_3$ растет быстрее, чем логарифм правдоподобия мультимодели $\mathbf{f}_1$. После 20-й итерации все мультимодели имеют одинаковое правдоподобие.

На фиг. 5–7 показан процесс сходимости для разных мультимоделей $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$. На фиг. 7 показана мультимодель $\mathbf{f}_1$, которая аппроксимирует окружности не верно. На фиг. 5, 6 показаны мультимодели $\mathbf{f}_2$, $\mathbf{f}_3$, которые аппроксимируют окружности верно.

Вычислительный эксперимент показывает, что мультимодели $\mathbf{f}_2$, $\mathbf{f}_3$ которые используют априорные распределения на параметры экспертов, аппроксимируют окружности лучше, чем мультимодель $\mathbf{f}_1$, которая работает без априорных распределений.

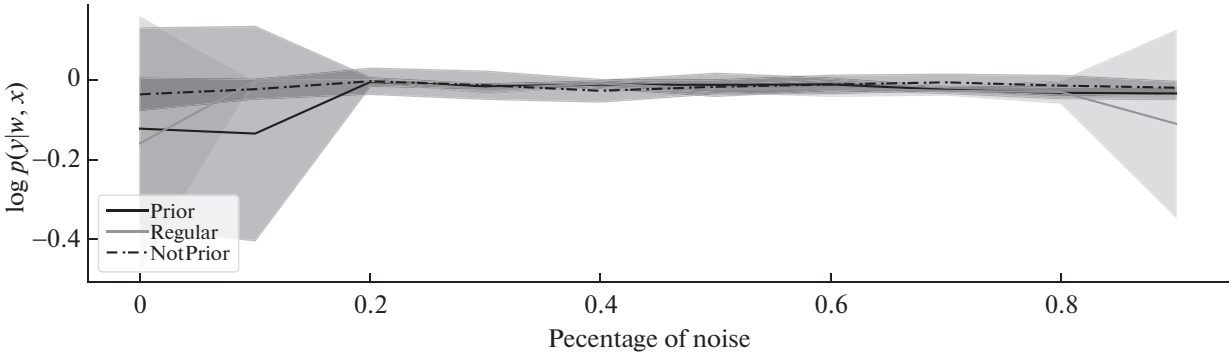


Фиг. 7. Визуализации процесса сходимости мультимодели без использования априорного распределения.

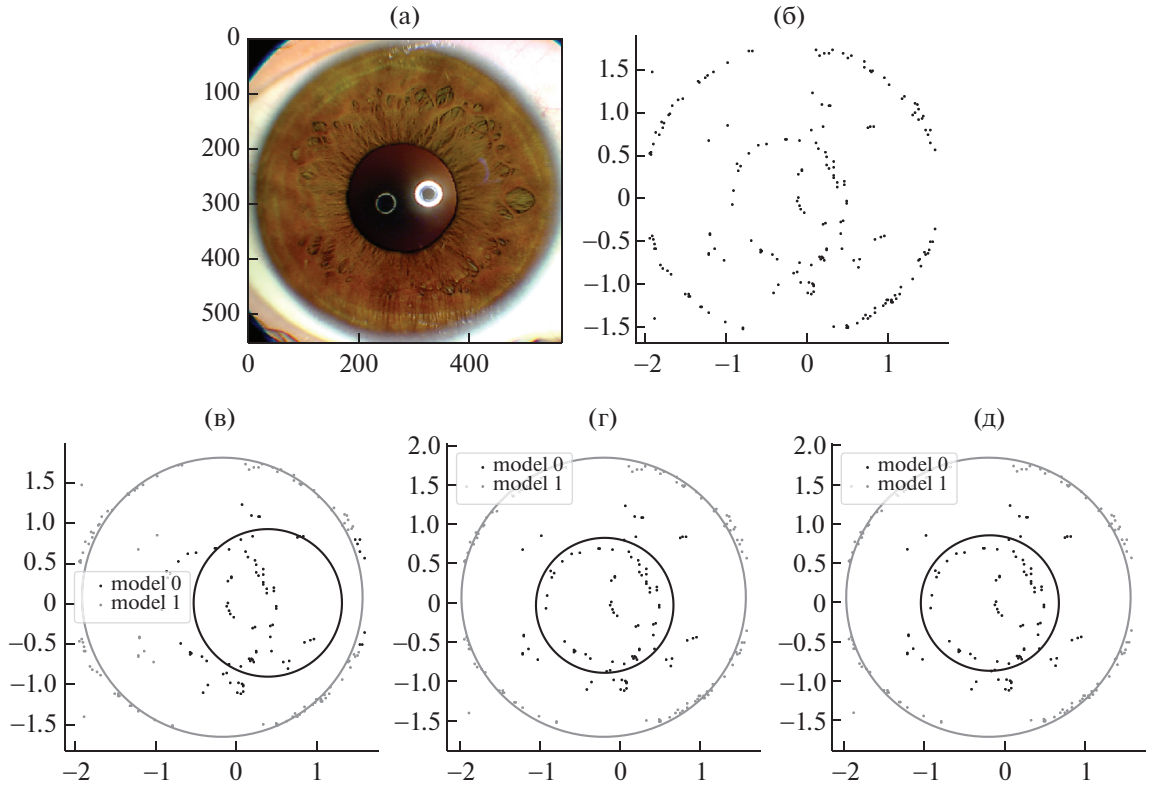


Фиг. 8. Зависимости центра и радиуса окружностей от номера итерации: (а), (б) – модель с регуляризацией априорных распределений; (в), (г) – модель с заданными априорными распределениями на параметры моделей; (д), (е) – модель без задания априорных распределений.

Анализ мультимоделей в зависимости от уровня шума. Данная часть эксперимента анализирует зависимость разных мультимоделей $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$ от уровня шума. Анализ всех мультимоделей проводится на выборке Synthetic 1 с добавлением разного уровня шума. Минимальный уровень шума

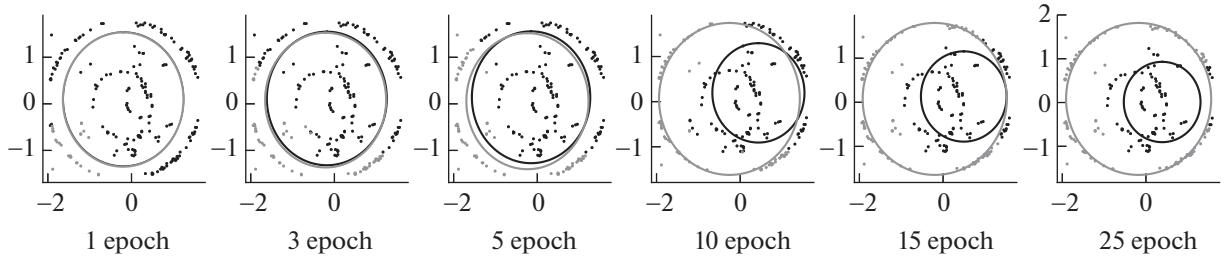


Фиг. 9. Зависимости логарифма правдоподобия (3.3) от уровня шума.

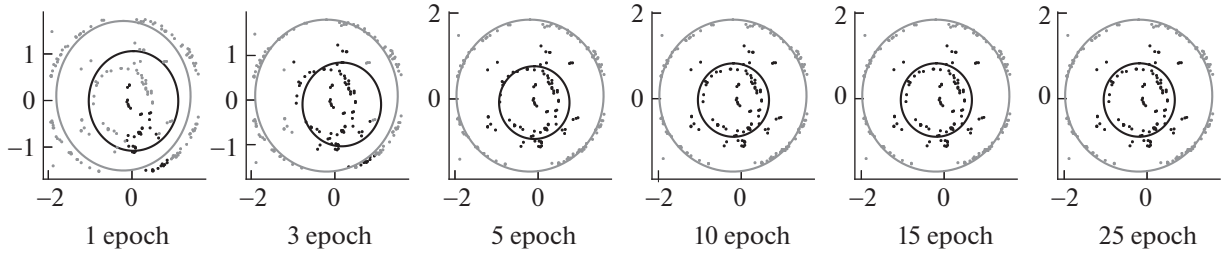


Фиг. 10. Мультимодель в зависимости от разных априорных предположений на реальном изображении: (а) – исходное изображение, (б) – бинаризованное изображение, (в) – мультимодель без априорных предположений, (г) – мультимодель с априорными распределениями на параметрах локальных моделей, (д) – мультимодель с регуляризацией на априорных распределениях параметров локальных моделей.

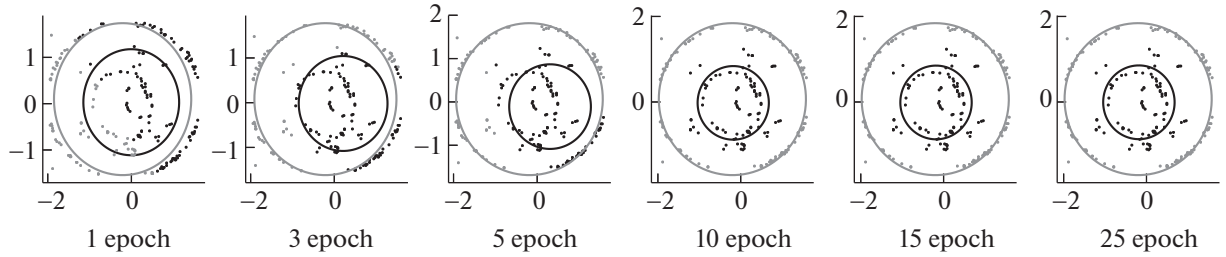
равен 0, когда число шумовых точек равно 0. Максимальный уровень шума равен 1, когда число шумовых точек равно числу точек на изображении. На фиг. 8 показаны график зависимости центра окружности и ее радиус в зависимости от уровня шума. Из графика следует, что радиус окружности увеличивается при увеличении уровня шума. Мультимодели $\mathbf{f}_2$, $\mathbf{f}_3$ аппроксимируют центр окружности верно, но мультимодель $\mathbf{f}_3$ более устойчива к шуму. На фиг. 9 показана зависимость логарифма правдоподобия (3.3) от уровня шума. Из графика следует, что логарифм правдоподобия (3.3) эквивалентный для всех мультимоделей, но на фиг. 8 видно, что качество аппроксимации (6.1) зависит от мультимодели. Данная часть вычислительного эксперимента показывает, что мультимодель $\mathbf{f}_3$ с регуляризацией априорного распределения является более устойчивой к шуму, чем остальные.



Фиг. 11. Визуализации процесса сходимости мультимодели без использования априорного распределения.



Фиг. 12. Визуализации процесса сходимости мультимодели с использованием априорного распределения.



Фиг. 13. Визуализации процесса сходимости мультимодели с использованием априорной регуляризации.

Реальные данные. Настоящая часть эксперимента анализирует разные мультимодели $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$ на реальной выборке. На фиг. 10 показан результат работы разных мультимоделей. Мультимодель $\mathbf{f}_1$ не верно аппроксимирует меньшую окружность. Мультимодели $\mathbf{f}_2$, $\mathbf{f}_3$ аппроксимируют обе окружности верно.

На фиг. 11–13 показан процесс аппроксимации для разных мультимоделей $\mathbf{f}_1$, $\mathbf{f}_2$, $\mathbf{f}_3$.

Данная часть эксперимента показывает, что мультимодели $\mathbf{f}_2$, $\mathbf{f}_3$ аппроксимируют окружности на реальных изображениях лучше, чем мультимодель $\mathbf{f}_1$.

7. ЗАКЛЮЧЕНИЕ

В настоящей работе сравниваются мультимодели, которые используют различные априорные предположения. Для анализа проводился вычислительный эксперимент на концентрических окружностях с разным уровнем шума. Для аппроксимации окружности на изображении использовалась линейная модель. Для взвешивания ответов разных линейных моделей использовалась шлюзовая функция, которая является двухслойным перцептроном с функцией softmax на последнем слое. В вычислительном эксперименте сравниваются мультимодели, которые используют априорное распределение и которые его не используют. Мультимодели, которые используют априорные распределения, имеют большую точность аппроксимации, чем мультимодель, которая не использует априорные распределения.

Также был проведен эксперимент по исследованию различных способов регуляризации априорных распределений параметров локальных моделей. В эксперименте показано, что в случае, когда регуляризация задана, мультимодель находит окружности более устойчиво. В эксперименте было показано, что все мультимодели являются чувствительными к выбросам. Для решения данной задачи предлагается использовать еще одну локальную модель, которая будет аппроксимировать шум.

В дальнейшем планируется улучшить мультимодель с помощью задания априорного распределения на шлюзовую функцию. Планируется рассмотреть в качестве моделей не только модели, которые описывают данные, но и модель, которая аппроксимирует шум в данных. Предполагается, что число шумовых точек мало, поэтому требуется задать априорное распределение, которое учитывает данную информацию.

СПИСОК ЛИТЕРАТУРЫ

1. *Tianqi C., Carlos G.* XGBoost: A Scalable Tree Boosting System // Proceed. 22nd ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining. 2016.
2. *Xi C., Hemant I.* Random Forests for Genomic Data Analysis // Genomics. 2012. Iss. 99. № 6. P. 323–329.
3. *Esen Y.S., Wilson J., Gader P.D.* Twenty Years of Mixture of Experts // IEEE Transact. Neural Networks and Learn. Syst. 2012. Iss. 23. № 8. P. 1177–1193.
4. *Rasmussen C.E., Ghahramani Z.* Infinite Mixtures of Gaussian Process Experts // Adv. Neural Informat. Proc. Syst. 14. 2002. P. 881–888.
5. *Shazeer N., Mirhoseini A., Maziarz K.* Outrageously large neural networks: the sparsely-gated mixture-of-experts layer // Internat. Conf. Learn. Representat. 2017.
6. *Jordan M.I.* Hierarchical mixtures of experts and the EM algorithm // Neural Comput. 1994. V. 6. № 2. P. 181–214.
7. *Jordan M.I., Jacobs R.A.* Hierarchies of adaptive experts // Adv. Neural Informat. Proc. Syst. 1991. P. 985–992.
8. *Lima C., Coelho A., Zuben F.J.* Hybridizing mixtures of experts with support vector machines: Investigation into nonlinear dynamic systems identification // Inf. Sci. 2007. V. 177. № 10. P. 2049–2074.
9. *Cao L.* Support vector machines experts for time series forecasting // Neurocomputing. 2003. V. 51. P. 321–339.
10. *Yumlu M.S., Gurgen F.S., Okay N.* Financial time series prediction using mixture of experts // Proc. 18th Int. Symp. Comput. Inf. Sci. 2003. P. 553–560.
11. *Cheung Y.M., Leung W.M., Xu L.* Application of mixture of experts model to financial time series forecasting // Proc. Int. Conf. Neural Netw. Signal Process. 1995. P. 1–4.
12. *Weigend A.S., Shi S.* Predicting daily probability distributions of S&P500 returns // J. Forecast. 2000. V. 19. № 4. P. 375–392.
13. *Ebrahimpour R., Moradian M.R., Esmkhani A., Jafarlou F.M.* Recognition of Persian handwritten digits using characterization loci and mixture of experts // J. Digital Content Technol. Appl. 2009. V. 3. № 3. P. 42–46.
14. *Estabrooks A., Japkowicz N.* A mixture-of-experts framework for text classification // Proc. Workshop Comput. Natural Lang. Learn., Assoc. Comput. Linguist. 2001. P. 1–8.
15. *Mossavat S., Amft O., Petkov Vries B., Kleijn W.* A Bayesian hierarchical mixture of experts approach to estimate speech quality // Proc. 2nd Int. Workshop Qual. Multimedia Exper. 2010. P. 200–205.
16. *Peng F., Jacobs R.A., Tanner M.A.* Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition // J. Amer. Stat. Assoc. 1996. V. 91. № 435. P. 953–960.
17. *Tuerk A.* The state based mixture of experts HMM with applications to the recognition of spontaneous speech. Ph.D. thesis. Cambridge: Univ. Cambridge, 2001.
18. *Sminchisescu C., Kanaujia A., Metaxas D.* Discriminative density propagation for visual tracking // IEEE Trans. Pattern Anal. Mach. Intell. 2007. V. 29. № 11. P. 2030–2044.
19. *Bowyer K., Hollingsworth K., Flynn P.* A Survey of Iris Biometrics Research: 2008–2010.
20. *Matveev I.* Detection of iris in image by interrelated maxima of brightness gradient projections // Appl. Comput. Math. 2010. V. 9. № 2. P. 252–257.
21. *Matveev I., Simonenko I.* Detecting precise iris boundaries by circular shortest path method // Pattern Recognit. and Image Anal. 2014. V. 24. P. 304–309.
22. *Dempster A.P., Laird N.M., Rubin D.B.* Maximum Likelihood from Incomplete Data via the EM Algorithm // J. the Royal Statist. Soc. Ser. B (Methodological). 1977. V. 39. № 1 P. 1–38.
23. *Bishop C.* Pattern Recognition and Machine Learning. Berlin: Springer, 2006. P. 758.