УДК 519.254

РОБАСТНАЯ ВЕРСИЯ ЕМ-АЛГОРИТМА ДЛЯ КОНЕЧНЫХ СМЕСЕЙ НОРМАЛЬНЫХ ЗАКОНОВ[1]

Горшенин А.К.[*], Казаков И.А.[**], Королев В.Ю.[***]
[*]Институт проблем информатики Российской академии наук, г. Москва
[**]Факультет ВМК, МГУ им. М.В. Ломоносова, г. Москва
[***]Факультет ВМК, МГУ им. М.В. Ломоносова; ИПИ РАН, г. Москва

Поступила в редакцию 16.08.2011, после переработки 20.08.2011.

В статье рассматривается устойчивая медианная модификация классического ЕМ-алгоритма для конечных смесей нормальных законов и приводится доказательство сходимости этого метода.

The paper considers stable median modification of the classical EM-algorithm for finite normal mixtures and gives a proof of the method's convergence.

Ключевые слова: робастность, медианные оценки, ЕМ-алгоритм, смеси нормальных распределений, смеси распределений Лапласа.
Keywords: robustness, median estimators, EM-algorithm, normal mixtures, mixtures of Laplace distributions.

Введение

Задача статистического оценивания параметров конечной смеси нормальных законов (иногда называемая задачей разделения смеси) возникает во многих приложениях, см., например, [1], где подробно описаны методы решения упомянутой задачи как средства декомпозиции волатильности финансовых индексов или анализа плазменной турбулентности. Самым популярным методом решения этой многопараметрической задачи является ЕМ-алгоритм. Впрочем, данный алгоритм обладает некоторыми критическими недостатками. Например, для случая четырехкомпонентной смеси нормальных законов при объеме выборки в 200−300 наблюдений замена лишь одного из них может кардинально изменить итоговые оценки ЕМ-алгоритма [1]. С целью преодоления неустойчивости ЕМ-алгоритма по исходным данным в работах [2, 3] были предложены робастные модификации ЕМ-алгоритма и его стохастической версии (SEM-алгоритма) для разделения конечных смесей нормальных законов. В этих модификациях для оценивания параметров компонент смеси вместо моментных оценок используются медианные оценки параметров сдвига и положения. Однако сходимость указанных медианных модификаций ЕМ-алгоритма в упомянутых работах не получила надлежащего теоретического обоснования. В частности, остались открытыми вопросы почему и к чему сходится

[1]Работа поддержана Российским фондом фундаментальных исследований (проект 11-01-12026-офи-м).

предложенный итерационный процесс. В данной статье предлагается обоснование сходимости медианных модификаций ЕМ-алгоритма.

1. ЕМ-алгоритм для разделения конечных смесей нормальных законов

Пусть дана реализация $\boldsymbol{x} = (x_1, \ldots, x_n)$ выборки $\boldsymbol{X} = (X_1, \ldots, X_n)$, где $X_1, \ldots, X_n$ – независимые случайные величины с одной и той же плотностью, имеющей вид смеси, то есть

$$\sum_{i=1}^{k} p_i \psi_i(x; t_i), \tag{1}$$

где $k \geqslant 1$ – известное натуральное число, $\psi_1, \ldots, \psi_k$ – известные плотности распределения, $p_i \geqslant 0$, $\sum_{i=1}^{k} p_i = 1$; t_i, $i = 1, \ldots, k$, – вообще говоря, многомерные параметры. Плотности $\psi_1, \ldots, \psi_k$ будем называть компонентами смеси (1), параметры $p_1, \ldots, p_k$ будем называть весами соответствующих компонент.

Задачей разделения смеси (1) принято называть задачу статистического оценивания параметров $p_1, \ldots, p_k, t_1, \ldots, t_k$ по известной реализации $\boldsymbol{x} = (x_1, \ldots, x_n)$ выборки $\boldsymbol{X} = (X_1, \ldots, X_n)$. Необходимым условием существования решения задачи разделения смеси вероятностных распределений вида (1) является идентифицируемость смеси.

Как уже отмечалось, для численного решения данной задачи чаще всего используется так называемый ЕМ-алгоритм. ЕМ-алгоритмом принято называть итерационную процедуру поиска оценок максимального правдоподобия вектора параметров $(p_1, \ldots, p_k, t_1, \ldots, t_k)$. Применительно к конечным смесям нормальных законов вида (1) ЕМ-алгоритм определяется следующим образом (см., например, [1]). В качестве плотностей $\psi_i(x; t_i)$ из формулы (1) рассмотрим

$$\psi_i(x; t_i) = \frac{1}{\sigma_i} \varphi\Big(\frac{x - a_i}{\sigma_i}\Big), \quad x \in \mathbb{R},$$

где

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\},$$

– плотность стандартного нормального распределения, $t_i = (a_i, \sigma_i^2)$, $a_i \in \mathbb{R}$, $\sigma_i > 0$, $i = 1, \ldots, k$. Обозначим

$$g_{ij}^{(m)} = \frac{\dfrac{p_i^{(m)}}{\sigma_i^{(m)}} \varphi\Big(\frac{x_j - a_i^{(m)}}{\sigma_i^{(m)}}\Big)}{\sum\limits_{r=1}^{k} \dfrac{p_r^{(m)}}{\sigma_r^{(m)}} \varphi\Big(\frac{x_j - a_r^{(m)}}{\sigma_r^{(m)}}\Big)} = \frac{\dfrac{p_i^{(m)}}{\sigma_i^{(m)}} \exp\Big\{ -\frac{1}{2}\Big(\frac{x_j - a_i^{(m)}}{\sigma_i^{(m)}}\Big)^2\Big\}}{\sum\limits_{r=1}^{k} \dfrac{p_r^{(m)}}{\sigma_r^{(m)}} \exp\Big\{ -\frac{1}{2}\Big(\frac{x_j - a_r^{(m)}}{\sigma_r^{(m)}}\Big)^2\Big\}}.$$

Здесь предполагается, что значения $(p_1^{(m)}, \ldots, p_k^{(m)}, a_1^{(m)}, \ldots, a_k^{(m)} \sigma_1^{(m)}, \ldots, \sigma_k^{(m)})$ параметров на m-й итерации ЕМ-алгоритма известны ($m \geq 0$). Величину $g_{ij}^{(m)}$ можно интерпретировать как статистическую оценку апостериорной вероятности того, что реализация x_j случайной величины X_j сгенерирована в соответствии с i-й компонентой смеси вида (1), то есть $g_{ij}^{(m)}$ является апостериорной вероятностью того, что распределением случайной величины x_j является $\Phi\big((x - a_i^{(m)})/\sigma_i^{(m)}\big)$, где $\Phi(x)$ – стандартная нормальная функция распределения, соответствующая плотности $\varphi(x)$. Тогда значения параметров p_i, a_i и σ_i на $(m+1)$-й итерации ЕМ-алгоритма соответственно определяются как

$$p_i^{(m+1)} = \frac{1}{n}\sum_{j=1}^{n} g_{ij}^{(m)} \tag{2}$$

$$a_i^{(m+1)} = \frac{1}{\sum_{j=1}^{n} g_{ij}^{(m)}}\sum_{j=1}^{n} g_{ij}^{(m)} x_j,$$

$$\sigma_i^{(m+1)} = \left[\frac{1}{\sum_{j=1}^{n} g_{ij}^{(m)}}\sum_{j=1}^{n} g_{ij}^{(m)} (x_j - a_i^{(m+1)})^2\right]^{1/2}, \ i = 1, \ldots, k.$$

Обратим внимание на то, что $a_i^{(m+1)}$ является выборочным средним, построенным по реализации $\boldsymbol{x} = (x_1, \ldots, x_n)$ выборки $\boldsymbol{X} = (X_1, \ldots, X_n)$, как если бы распределение каждого ее элемента задавалось вероятностями

$$p_{ij}^{(m)} = \frac{g_{ij}^{(m)}}{\sum_{j=1}^{n} g_{ij}^{(m)}}, \quad i = 1, \ldots, k, \ j = 1, \ldots, n. \tag{3}$$

Для борьбы с неустойчивостью классического ЕМ-алгоритма по отношению к исходным данным можно использовать робастные оценки на шагах ЕМ-алгоритма, то есть оценки, обладающие нечувствительностью к малым отклонениям от предположений (подробнее об этом см. книгу П. Хьюбера [4]). В качестве робастных оценок в вышеупомянутой книге [4] предложено использовать так называемые М-оценки. М-оценка – всякая оценка T_n, определяемая как решение экстремальной задачи на минимум вида

$$\sum_{i=1}^{n} \rho(x_i; T_n) \to \min,$$

где $\rho(\cdot)$ – произвольная функция. Заметим, что если в качестве $\rho(x;\theta)$ взять функцию $-\log f(x;\theta)$, где $f(x;\theta)$ – плотность распределения наблюдений, θ – неизвестный параметр, то можно получить оценки максимального правдоподобия. Данные оценки допускают обобщение на многопараметрический случай, что позволяет одновременно выписывать оценки данного типа для сдвига и масштаба.

В книге [4] показано, что медиана является робастной М-оценкой параметра сдвига. Более того, известно, что медиана является единственной М-оценкой, инвариантной относительно масштаба. Поэтому в данной работе значительное внимание уделяется построению именно медианной модификации ЕМ-алгоритма для конечных смесей нормальных законов.

2. Медианная модификация ЕМ-алгоритма

Как было экспериментально установлено, ЕМ-алгоритм обладает сильной неустойчивостью по исходным данным. Как уже отмечалось, в случае четырехкомпонентной смеси нормальных законов при объеме выборки $200 - 300$ наблюдений замена лишь одного наблюдения другим может кардинально изменить итоговые оценки, полученные с помощью ЕМ-алгоритма.

Для противодействия указанной неустойчивости ЕМ-алгоритма можно использовать его медианные модификации. Смысл этих модификаций в том, что наиболее неустойчивые этапы выполнения ЕМ-алгоритма заменяются более устойчивыми. В частности, на М-этапе неустойчивые моментные оценки наибольшего правдоподобия заменяются более устойчивыми (робастными) оценками медианного типа. Опишем возможный вариант подобной модификации.

Пусть числа $g_{ij}^{(m)}$ известны. По числам $g_{ij}^{(m)}$ определим «вероятности» $p_{ij}^{(m)}$ по правилу

$$p_{ij}^{(m)} = g_{ij}^{(m)} \left(\sum_{j=1}^{n} g_{ij}^{(m)} \right)^{-1}, \quad i = 1, \ldots, k; \; j = 1, \ldots, n$$

(n – объем выборки, k – число компонент смеси). Через $\boldsymbol{x} = (x_1, \ldots, x_n)$ обозначим выборку. Тогда число $p_{ij}^{(m)}$ можно интерпретировать как вероятность того, что наблюдение x_j имеет распределение, определяемое i-й компонентой смеси.

Введем «фиктивные» случайные величины $\xi_i^{(m)}$, $i = 1, \ldots, k$, которые соответственно принимают значение x_j с вероятностями $p_{ij}^{(m)}$, $i = 1, \ldots, k$, $j = 1, \ldots, n$ (несложно видеть, что $\sum_{j=1}^{n} p_{ij}^{(m)} = 1$). При этом оценка параметра сдвига i-й компоненты смеси на $(m+1)$-й итерации оказывается в точности равной математическому ожиданию случайной величины $\xi_i^{(m)}$:

$$a_i^{(m+1)} = \frac{1}{\sum_{j=1}^{n} g_{ij}^{(m)}} \sum_{j=1}^{n} g_{ij}^{(m)} x_j = \sum_{j=1}^{n} p_{ij}^{(m)} x_j = \mathbb{E}_{\theta^{(m)}} \xi_i^{(m)}.$$

Для того чтобы построить модификацию ЕМ-алгоритма, более устойчивую по отношению к наличию «засоряющих» наблюдений (а при оценивании параметров какой-либо компоненты смеси наблюдения, распределения которых соответствуют другим компонентам, неизбежно будут «засоряющими» по отношению к оцениваемой компоненте), в качестве оценки параметра a_i на $(m+1)$-й итерации предлагается взять медиану $\operatorname{med} \xi_i^{(m)}$ случайной величины $\xi_i^{(m)}$, которую можно вычислить так. Переупорядочим значения $x_1, \ldots, x_n$ случайной величины $\xi_i^{(m)}$ по неубыванию. Получим вариационный ряд $x_{(1)}, \ldots, x_{(n)}$. Ясно, что одно и то же переупорядочение имеет место для значений всех случайных величин $\xi_i^{(m)}$. Одновременно переставятся и вероятности $p_{ij}^{(m)}$, соответствующие значениям каждой случайной величины $\xi_i^{(m)}$. Пусть $\widehat{p}_{ij}^{(m)}$ – это та из вероятностей $p_{ij}^{(m)}$, которая соответствует значению $x_{(j)}$ случайной величины $\xi_i^{(m)}$. Положим

$$J_i = \min\{j : \widehat{p}_{i1}^{(m)} + \widehat{p}_{i2}^{(m)} + \ldots + \widehat{p}_{ij}^{(m)} \geq \tfrac{1}{2}\}.$$

Тогда

$$a_i^{(m+1)} = \operatorname{med} \xi_i^{(m)} = x_{(J_i)}. \tag{4}$$

Вычислим $\mathbb{E}_{\theta^{(m+1)}}\left|\zeta_i^{(m+1)} - a_i^{(m+1)}\right|$. Имеем

$$\mathbb{E}_{\theta^{(m+1)}}\left|\zeta_i^{(m+1)} - a_i^{(m+1)}\right| = \int\limits_{-\infty}^{\infty} \left|x - a_i^{(m+1)}\right| d_x\Phi\left(\frac{x - a_i^{(m+1)}}{\sigma_i^{(m+1)}}\right) =$$

$$= 2\int\limits_{0}^{\infty} x d_x\Phi\left(\frac{x}{\sigma_i^{(m+1)}}\right) = \sigma_i^{(m+1)}\sqrt{\frac{2}{\pi}}.$$

Эмпирическим аналогом величины $\mathbb{E}_{\theta^{(m+1)}}\left|\zeta_i^{(m+1)} - a_i^{(m+1)}\right|$ является величина

$$s_i^{(m+1)} = \mathbb{E}_{\theta^{(m)}}\left|\xi_i^{(m)} - a_i^{(m+1)}\right| = \sum_{j=1}^{n} p_{ij}^{(m)}\left|x_j - a_i^{(m+1)}\right|.$$

Реализуя метод моментов и приравнивая величину $\mathbb{E}_{\theta^{(m+1)}}\left|\zeta_i^{(m+1)} - a_i^{(m+1)}\right|$ ее эмпирическому аналогу, получаем оценку для параметра σ_i на $(m+1)$-ой итерации:

$$\sigma_i^{(m+1)} = \sqrt{\frac{\pi}{2}} \cdot s_i^{(m+1)}. \tag{5}$$

Оценки $p_i^{(m+1)}$ весов p_i в модели (1) ищутся, как и ранее, по формулам (2). Числа же $g_{ij}^{(m+1)}$ на каждой итерации переназначаются по следующему правилу:

$$g_{ij}^{(m+1)} = \frac{\dfrac{p_i^{(m+1)}}{\sigma_i^{(m+1)}} \exp\left\{-\dfrac{1}{2}\left(\dfrac{x_j - a_i^{(m+1)}}{\sigma_i^{(m+1)}}\right)^2\right\}}{\sum\limits_{r=1}^{k} \dfrac{p_r^{(m+1)}}{\sigma_r^{(m+1)}} \exp\left\{-\dfrac{1}{2}\left(\dfrac{x_j - a_r^{(m+1)}}{\sigma_r^{(m+1)}}\right)^2\right\}}. \tag{6}$$

Таким образом, медианная модификация ЕМ-алгоритма определяется соотношениями (2), (4), (5) и (6).

3. Обоснование целесообразности применения медианной модификации ЕМ-алгоритма

Можно показать (см., например, книгу [1]), что для максимизации параметров в процессе итерационных шагов ЕМ-алгоритма, необходимо максимизировать функцию вида

$$\sum_{i=1}^{k}\sum_{j=1}^{n} g_{ij}^{(m)} \log \frac{\psi\left(\frac{x_j - a_i}{\sigma_i}\right)}{\sigma_i}.$$

где величины $g_{ij}^{(m)}$ имеют вид

$$g_{ij}^{(m)} = \frac{\dfrac{p_i^{(m)}}{\sigma_i}\psi\left(\frac{x_j - a_i^{(m)}}{\sigma_i^{(m)}}\right)}{\sum\limits_{r=1}^{k} \dfrac{p_r^{(m)}}{\sigma_r}\psi\left(\frac{x_j - a_r^{(m)}}{\sigma_r^{(m)}}\right)},$$

а величины $\psi(\cdot)$ обозначают плотности некоторых распределений. Рассмотрим в качестве примера распределение Лапласа, задаваемое плотностью вида

$$l(x) = \frac{1}{\sqrt{2}} \exp\left\{-\sqrt{2}|x|\right\}. \tag{7}$$

Тогда

$$\sum_{i=1}^{k}\sum_{j=1}^{n} g_{ij}^{(m)} \log \frac{\psi\left(\frac{x_j-a_i}{\sigma_i}\right)}{\sigma_i} = \sum_{i,j} g_{ij}^{(m)} \cdot \left(-\frac{1}{2}\log 2 - \log\sigma_i - \sqrt{2}\frac{|x_j-a_i|}{\sigma_i}\right) =$$
$$= -\frac{1}{2}\log 2 \sum_{i=1}^{k}\sum_{j=1}^{n} g_{ij}^{(m)} - \sum_{i=1}^{k}\log\sigma_i \sum_{j=1}^{n} g_{ij}^{(m)} - \sqrt{2}\sum_{i=1}^{k}\sum_{j=1}^{n} g_{ij}^{(m)}\frac{|x_j-a_i|}{\sigma_i}. \tag{8}$$

В данной сумме первое слагаемое не зависит от параметров a_i и σ_i, поэтому для отыскания значений a_i и σ_i, максимизирующих все выражение, необходимо минимизировать второе и третье слагаемые. При этом с учетом выражения (3) третье слагаемое имеет вид

$$\sqrt{2}\sum_{i=1}^{k}\sum_{j=1}^{n} g_{ij}^{(m)}\frac{|x_j-a_i|}{\sigma_i} = \sqrt{2}\sum_{i=1}^{k}\frac{1}{\sigma_i}\left(\sum_{j=1}^{n} g_{ij}^{(m)}\right)\sum_{j=1}^{n} p_{ij}^{(m)}|x_j-a_i| =$$
$$= \sqrt{2}\sum_{i=1}^{k}\frac{1}{\sigma_i}\left(\sum_{j=1}^{n} g_{ij}^{(m)}\right)\mathbb{E}|\xi_i^{(m)}-a_i|.$$

Здесь случайные величины $\xi_i^{(m)}$ принимают значение x_j с вероятностями $p_{ij}^{(m)}$, определяемыми формулой (3), $i = 1,\ldots,k$, $j = 1,\ldots,n$ (несложно видеть, что $\sum_{j=1}^{n} p_{ij}^{(m)} = 1$). Используя хорошо известное свойство медианы, замечаем, что при каждом $i = 1,\ldots,k$

$$\arg\min_a \mathbb{E}|\xi_i^{(m)} - a| = \operatorname{med}\xi_i^{(m)}.$$

Отсюда вытекает, что для конечной смеси распределений Лапласа оптимальная оценка параметра a_i на $(m+1)$-й итерации EM-алгоритма имеет вид

$$a_i^{(m+1)} = \operatorname{med}\xi_i^{(m)}. \tag{9}$$

Подставляя это значение $a_i^{(m+1)}$ в (8) и максимизируя полученное выражение по σ_i путем дифференцирования по σ_i и приравнивания производной нулю, мы приходим к оценке

$$\sigma_i^{(m+1)} = \sqrt{2}\sum_{j=1}^{n} p_{ij}^{(m)}|x_j - a_i^{(m+1)}|.. \tag{10}$$

С целью дальнейшего исследования свойств последовательностей, определяемых соотношениями (2), (9), (10), наряду с «базовой» смесью

$$\sum_{i=1}^{k} p_i\Phi\left(\frac{x-a_i}{\sigma_i}\right) \tag{11}$$

рассмотрим смесь

$$\sum_{i=1}^{k} p_i L\left(\frac{x-a_i}{\sigma_i}\right), \tag{12}$$

где $L(\cdot)$ – функция распределения Лапласа (двойного экспоненциального распределения), соответствующая плотности (7),

$$L(x) = \begin{cases} \frac{1}{2}e^{\sqrt{2}x}, & x \le 0, \\ 1 - \frac{1}{2}e^{-\sqrt{2}x}, & x > 0. \end{cases}$$

Ядра обеих смесей симметричны, а их центры – параметры положения a_i компонент – совпадают (так как математическое ожидание каждой компоненты каждой из смесей (11) и (12) равно a_i). Последовательность оценок, определяемая соотношениями (2), (9), (10), является последовательностью оценок *классического ЕМ-алгоритма* для смеси вида (12) и, стало быть, при надлежащем выборе начального приближения и критерия остановки сходится к оценкам максимального правдоподобия параметров модели (12). Поскольку параметры положения компонент моделей (11) и (12) совпадают, то оценки этих параметров в рамках модели (12) (и в определенном смысле оптимальные для этой модели) будут и оценками тех же параметров положения компонент в рамках модели (11) (но не обязательно оптимальными для этой модели). Но в силу причин, описанных выше, они будут устойчивыми. Так как оценки параметров масштаба компонент в рамках обеих моделей совпадают с точностью до неслучайного коэффициента (для описанного выше медианного ЕМ-алгоритма он равен $\sqrt{\pi/2}$), то аналогичное заключение справедливо и для оценок параметров положения. Более того, алгоритм, определяемый соотношениями (2), (9), (10) и являющийся монотонным в рамках модели (12), останавливается за конечное число итераций при любом разумном критерии остановки.

Таким образом, медианные оценки параметров положения компонент естественным образом возникают на Е-этапе классического ЕМ-алгоритма в задаче разделения конечных смесей двойных экспоненциальных распределений (распределений Лапласа) (12) с теми же самыми значениями параметров сдвига и масштаба компонент, что и у исходной смеси нормальных законов (11). Значит, по сути медианная модификация ЕМ-алгоритма сводится к замене исходной задачи разделения конечных смесей нормальных законов задачей разделения конечных смесей распределений Лапласа с теми же самыми значениями параметров сдвига и масштаба компонент.

Для интерпретации и обоснования возможности подобной замены заметим, что двойное экспоненциальное распределение можно представить в виде масштабной смеси нормальных законов при стандартном показательном смешивающем распределении. Действительно, легко убедиться, что, если U – случайная величина с функцией распределения $\mathbb{P}(U < x) = (1 - e^{-x})\mathcal{I}(x \ge 0)$, то (см., например, [1])

$$\mathbb{E}\Phi\left(\frac{x}{\sqrt{U}}\right) = L(x).$$

Поэтому при указанной замене исходные данные представляются в виде «зашумленной» выборки, причем «зашумление» производится с помощью умножения параметров масштаба компонент на случайную величину со стандартным показательным распределением, а подлежащие оцениванию параметры положения (сдвига) компонент остаются неизменными. Искусственно «зашумляя» результаты экспоненциальным распределением, которое, как известно, является наиболее энтропийным среди всех распределений, сосредоточенных на положительной оси и имеющих такое же математическое ожидание, мы, фактически, можем избавиться от лишних деталей, которые подвержены сильному изменению при малых объемах выборки, а значит, уменьшить негативное влияние неустойчивости на результаты работы алгоритма. При этом еще раз отметим, что указанное «зашумление» не искажает параметры положения компонент. Для большей наглядности можно предложить следующую аналогию. Представим себе, что в контрастной фотографии, содержащей много деталей, нужно выделить несколько «главных» по интенсивности точек. Автоматический поиск таких точек затруднен обилием локальных максимумов интенсивности за счет большого числа мелких деталей. Но если посмотреть на фотографию через матовое стекло («зашумить» изображение), то мелкие детали сольются, а различимыми останутся только «главные» по интенсивности места исходного изображения, которые и будут легко обнаружены автоматической процедурой поиска таких мест.

Таким образом, робастная версия классического EM-алгоритма, основанная на медианных оценках, может быть определена формулами (2), (9) и (10) (с точностью до неслучайного множителя), то есть полностью соответствует рассмотренному в разделе медианному EM-алгоритму, определяемому формулами (2), (4) и (5).

Заключение

Итак, было показано, что медианные оценки естественным образом возникают на E-этапе EM-алгоритма в задаче разделения конечных смесей двойных экспоненциальных распределений с теми же самыми значениями параметров сдвига и масштаба компонент, что и у исходной смеси нормальных законов. В свою очередь, двойное экспоненциальное распределение можно представить в виде масштабной смеси нормальных законов при стандартном показательном смешивающем распределении. Таким образом, медианная модификация EM-алгоритма по сути сводится к замене исходной задачи разделения конечных смесей нормальных законов задачей разделения конечных смесей распределений Лапласа с теми же самыми значениями параметров сдвига и масштаба компонент. При этом оценки, получаемые с помощью медианной версии EM-алгоритма в задаче разделения конечных смесей нормальных законов, приближают оцениваемые параметры постольку, поскольку соответствующая последовательность оценок, получаемая EM-алгоритмом, сходится к оценкам максимального правдоподобия аналогичных параметров в модели вида конечных смесей распределения Лапласа.

Подводя итог, можно заключить, что использование робастных медианных оценок параметров компонент конечных смесей нормальных законов на каждом шаге EM-алгоритма довольно разумно и может давать более «четкие» результирующие оценки параметров конечных смесей нормальных законов. Данные выводы подтверждаются на практике при применении медианной модификации EM-алгоритма к реальным данным – финансовым индексам, турбулентной плазме (см., например, работы [2, 3, 5]).

Список литературы

[1] Королев В. Ю. Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. – М.: изд-во Моск. ун-та, 2011. – 512 с.

[2] Горшенин А. К., Королев В. Ю., Турсунбаев А. М. Медианные модификации EM-алгоритма для разделения смесей вероятностных распределений и их применение к декомпозиции волатильности финансовых индексов // Статистические методы оценивания и проверки гипотез. Пермь, изд-во Пермского государственного университета, 2008. – С. 169–195.

[3] Горшенин А. К., Королев В. Ю., Турсунбаев А. М. Медианные модификации EM- и SEM-алгоритмов для разделения смесей вероятностных распределений и их применение к декомпозиции волатильности финансовых временных рядов // Информатика и ее применения, 2008. Т. 2. Вып. 4. С. 12–47.

[4] Хьюбер П. Робастность в статистике. М.: Мир,1984. – 304 с.

[5] Батанов Г. М, Горшенин А. К., Королев В. Ю., Малахов Д. В., Скворцова Н. Н. Эволюция вероятностных характеристик низкочастотной турбулентности плазмы в микроволновом поле // Математическое моделирование, 2011. Т. 23. № 5. С. 35–55.