УДК 519.217.2

О СХОДИМОСТИ ПОСЛЕДОВАТЕЛЬНОСТИ SEM-ОЦЕНОК В ЗАДАЧЕ СТАТИСТИЧЕСКОГО РАЗДЕЛЕНИЯ СМЕСЕЙ [1]

Горшенин А.К.
Институт проблем информатики Российской академии наук, г. Москва

Поступила в редакцию 23.09.2011, после переработки 23.11.2011.

В работе доказана теорема об асимптотических свойствах последовательности оценок параметров смеси вероятностных распределений, получаемой SEM-алгоритмом, в задаче статистического разделения произвольных конечных идентифицируемых смесей вероятностных распределений без дополнительных ограничений.

In the paper the theorem about the asymptotic properties of the SEM-algorithm parameter estimates sequence for the problem of statistical separation of arbitrary finite identifiable mixtures of probability distributions is proved without any restrictions.

Ключевые слова: смеси вероятностных распределений, статистическое разделение смесей, SEM-алгоритм, цепь Маркова, эргодичность.
Keywords: mixture distributions, statistical separation of mixtures, SEM-algorithm, Markov chain, ergodicity.

Введение

Для анализа стохастической структуры хаотической системы часто используются модели типа конечной смеси вероятностных распределений, а сама задача анализа сводится к статистическому разделению конечных смесей, то есть к отысканию статистических оценок параметров смеси.

Подобные модели находят эффективное применение в целом ряде отраслей: финансовом секторе (изучение скрытых тенденций эволюции различных секторов рынка или различных финансовых инструментов, основанное на применении понятия многомерной волатильности), физике турбулентной плазмы (анализ распределения энергии между процессами или структурами, исследование корреляционной структуры хаотических процессов), информационных системах (исследование стохастической структуры информационных потоков в вычислительных или телекоммуникационных системах) и так далее.

[1] Работа поддержана Российским фондом фундаментальных исследований (проекты 11-07-00112а, 11-01-12026-офи-м), а также Министерством образования и науки РФ в рамках ФЦП «Научные и научно-педагогические кадры инновационной России на 2009-2013 годы».

Для решения задачи статистического разделения смесей используются различные методы, наиболее популярным из которых является EM-алгоритм. EM-алгоритм представляет собой итеративный метод для нахождения оценок максимального правдоподобия. Однако данный алгоритм обладает некоторыми критическими недостатками. Например, классический EM-алгоритм выбирает ближайший к начальному приближению локальный максимум, который не обязательно является глобальным максимумом функции правдоподобия. Для борьбы с данным недостатком предлагается использовать стохастическую модификацию EM-алгоритма (так называемый SEM-алгоритм [1]). Подробное изучение свойств SEM-алгоритма можно найти, например, в работах [2, 3, 4]. Однако в указанных работах при доказательстве свойств сходимости используются дополнительные ограничения, которые могут не выполняться при практическом применении SEM-алгоритма. В настоящей работе доказывается теорема о свойствах сходимости последовательности оценок SEM-алгоритма без дополнительных предположений для произвольных семейств распределений.

1. Стохастическая модификация EM-алгоритма

Предположим, что плотность наблюдаемой случайной величины X имеет вид

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

где $k \geqslant 1$ – известное натуральное число, $\psi_1, \ldots, \psi_k$ – известные плотности распределения, неизвестный параметр θ имеет вид

$$\theta = (p_1, \ldots, p_k; t_1, \ldots, t_k),$$

причем $p_i \geqslant 0$, $\sum\limits_{i=1}^{k} p_i = 1$; 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 θ по известным реализациям случайной величины X. Необходимым условием существования решения задачи разделения смеси вероятностных распределений вида (1) является идентифицируемость смеси (см. [1]).

Предположим, что в нашем распоряжении имеется независимая выборка значений $\mathbf{x} = (x_1, \ldots, x_n)$ наблюдаемой случайной величины X, относительно которой, не ограничивая общности, будем предполагать, что ее распределение абсолютно непрерывно относительно меры Лебега (откуда необходимо вытекает, что $\psi_1, \ldots, \psi_k$ – это также плотности относительно меры Лебега). Также будем предполагать, что все x_j выбираются из множества $\mathfrak{D} = \{\psi_i(x; t_i) > 0,\ i = 1, \ldots, k\}$.

В рамках модели (1) логарифм классической (неполной) функций правдоподобия параметра θ имеет вид

$$\log L(\theta; \mathbf{x}) = \log \prod_{j=1}^{n} f_\theta^X(x_j) = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{k} p_i \psi_i(x_j; t_i) \right).$$

Непосредственный поиск точки максимума этой функции весьма затруднителен. Однако, если трактовать наблюдения $\mathbf{x}$ как неполные, то функцию правдоподобия можно записать в намного более удобном виде.

Предположим, что наряду с наблюдаемой случайной величиной X задана ненаблюдаемая случайная величина Y, значения которой содержат информацию о номерах компонент, в соответствии с которыми «генерируются» наблюдения $\mathbf{x} = (x_1, \ldots, x_n)$. А именно, будем считать, что наблюдения организованы следующим образом. При очередном, скажем, j-ом наблюдении $(j = 1, \ldots, n)$ сначала реализуется значение $y_j \in \{1, \ldots, k\}$ ненаблюдаемой случайной величины Y. Это значение y_j имеет смысл номера той компоненты смеси, которая затем выбирается в качестве распределения наблюдаемой случайной величины X при j-ом измерении, результатом которого является значение x_j. Такая схема типична для задач кластерного анализа, в которых каждое наблюдение может быть порождено одной и только одной компонентой смеси. Эта схема оказывается формально очень удобной для решения статистической задачи разделения конечных смесей вида (1).

Чтобы описать SEM-алгоритм, представим ненаблюдаемую информацию в следующей форме. Будем считать, что каждому наблюдению x_j соответствует вектор $\vec{y}_j = (y_{1j}, y_{2j}, \ldots, y_{kj})^T$, $j = 1, \ldots, n$, где k – число компонент смеси, n – объем выборки. При этом

$$y_{ij} = \begin{cases} 1, & \text{если наблюдение } x_j \text{ порождено } i\text{-й компонентой смеси,} \\ 0, & \text{в противном случае.} \end{cases}$$

При каждом j единице равна только одна из компонент вектора $\vec{y}_j$, остальные компоненты этого вектора равны нулю.

В терминах величин $\mathbf{y} = \left\{\vec{y}_j = (y_{1j}, y_{2j}, \ldots, y_{kj})^T,\ j = 1, \ldots, n\right\}$ логарифм полной функции правдоподобия для модели (1) принимает вид

$$\log L(\theta; x, y) = \sum_{j=1}^{n} \sum_{i=1}^{k} y_{ij} \log[p_i \psi_i(x_j; t_i)] =$$

$$= \sum_{i=1}^{k} \log p_i \sum_{j=1}^{n} y_{ij} + \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij} \log \psi_i(x_j; t_i). \qquad (2)$$

Векторы $\vec{y}_j = (y_{1j}, y_{2j}, \ldots, y_{kj})^T$, $j = 1, \ldots, n$, разбивают исходную наблюдаемую выборку $\mathbf{x}$ на k классов (кластеров) $K_1, \ldots, K_k$:

$$\mathbf{x} = K_1 \cup \ldots \cup K_k.$$

Для каждого $i = 1, \ldots, k$ с формальной точки зрения K_i – это множество тех наблюдений x_j, каждому из которых соответствует $y_{ij} = 1$. При этом каждое наблюдение x_j входит ровно в один кластер, то есть $K_i \cap K_j = \varnothing$ при $i \neq j$. Пусть υ_i – это число наблюдений, попавших в кластер K_i, $i = 1, \ldots, k$,

$$\upsilon_i = \sum_{j=1}^{n} y_{ij}.$$

Очевидно, что $\upsilon_1+\ldots+\upsilon_k = n$. Тогда, продолжая равенство (2), для логарифма полной функции правдоподобия в модели (1) получаем представление

$$\log L(\theta; x, y) = \sum_{i=1}^{k} \upsilon_i \log p_i \sum_{i=1}^{k} \sum_{j:\ x_j \in K_i} \log \psi_i(x_j; t_i). \tag{3}$$

Если бы величины y_{ij} были известны, то искать значение θ, максимизирующее функцию правдоподобия в выражении (3), можно было бы, максимизируя по θ каждое из слагаемых в правой части равенства (3), поскольку эти слагаемые зависят только от «своих» групп параметров. С помощью метода неопределенных множителей Лагранжа несложно убедиться, что максимум первого слагаемого по набору $p_1, \ldots, p_k$ при очевидном ограничении $p_1 + \ldots + p_k = 1$ достигается при

$$p_i^* = \frac{\upsilon_i}{n}. \tag{4}$$

Далее заметим, что

$$\sum_{j:\ x_j \in K_i} \log \psi_i(x_j; t_i) = \log \prod_{j:\ x_j \in K_i} \psi_i(x_j; t_i) \equiv \log L_i(t_i; K_i),$$

где $L_i(t_i; K_i)$ – это функция правдоподобия параметра t_i, построенная по подвыборке (кластеру) K_i в предположении, что каждый элемент подвыборки имеет плотность распределения $\psi_i(x_j; t_i)$. Отсюда видно, что, значения

$$t_i^* = \arg\max L_i(t_i; K_i),\ i = 1, \ldots, k, \tag{5}$$

доставляют максимум второму слагаемому в правой части равенства (3). Легко видеть, что соотношение (5) определяет обычные оценки наибольшего правдоподобия для параметров i-й компоненты смеси (1), построенные по подвыборке наблюдений, распределение которых равно этой компоненте, то есть по кластеру K_i.

Таким образом, если бы величины y_{ij} были известны, то оценки наибольшего правдоподобия параметров модели (1) определялись бы соотношениями (4) и (5). Однако на практике величины y_{ij} неизвестны. Идея SEM-алгоритма заключается в том, что эти величины определяются с помощью специального имитационного моделирования.

Итерационный SEM-алгоритм определяется так. Предположим, что известны значения $g_{ij}^{(m)}$ апостериорных вероятностей принадлежности наблюдения x_j к кластеру K_i, $i = 1, \ldots, k$; $j = 1, \ldots, n$; m – номер итерации (отметим, что

$$\sum_{i=1}^{k} g_{ij}^{(m)} = 1 \tag{6}$$

для каждого j и при каждом m).

На первом этапе SEM-алгоритма (*S-этапе*, от слов *Stochastic* или *Simulation*) для каждого $j = 1, \ldots, n$ генерируются векторы $\vec{y}_j^{(m+1)} =$

$(y_{1j}^{(m+1)}, y_{2j}^{(m+1)}, \ldots, y_{kj}^{(m+1)})^T$ как реализации случайных векторов с полиномиальным распределением с параметрами 1 и $g_{1j}^{(m)}, \ldots, g_{kj}^{(m)}$ ($g_{ij}^{(m)}$ – это вероятность того, что величина $y_{ij}^{(m+1)}$ равна единице). По векторам $\vec{y}_j^{(m+1)}$ определяется разбиение выборки $\mathbf{x} = (x_1, \ldots, x_n)$ на кластеры $K_1^{(m+1)}, \ldots, K_k^{(m+1)}$ и соответствующие числа $\upsilon_1^{(m+1)}, \ldots, \upsilon_k^{(m+1)}$ (численности кластеров) на $(m+1)$-й итерации. На S-этапе реализуется случайное «встряхивание» исходной выборки.

На втором этапе (*M-этапе*), этапе *максимизации*, в соответствии с формулами (3) и (4) вычисляются оценки максимального правдоподобия компонент параметра θ:

$$p_i^{(m+1)} = \frac{\upsilon_i^{(m+1)}}{n}, \tag{7}$$
$$t_i^{(m+1)} = \arg\max_t L_i(t; K_i^{(m+1)}), i = 1, \ldots, k.$$

Наконец, на третьем этапе (*E-шаге*) переназначаются вероятности g_{ij}. Название этого этапа восходит к слову *expectation*. Это обусловлено тем, что если $\vec{Y}_j^{(m+1)} = (Y_{1j}^{(m+1)}, Y_{2j}^{(m+1)}, \ldots, Y_{kj}^{(m+1)})^T$ – это случайный вектор, реализацией которого является вектор $\vec{y}_j^{(m+1)}$, а $\vec{X} = (X_1, \ldots, X_n)$ – это случайный вектор, реализацией которого является выборка $\mathbf{x} = (x_1, \ldots, x_n)$, то по определению

$$g_{ij}^{(m+1)} = \mathbb{E}_{\theta^{(m+1)}}(Y_{ij}^{(m+1)} \mid X_j) \tag{8}$$

(так как $Y_{ij}^{(m+1)}$ – это индикатор (случайного) события $\left\{X_j \in K_i^{(m+1)}\right\}$, а математическое ожидание индикатора случайного события равно вероятности этого события). При известном значении $X_j = x_j$ имеем

$$g_{ij}^{(m+1)} = \frac{p_i^{(m+1)} \psi_i(x_j; t_i^{(m+1)})}{\sum\limits_{r=1}^{k} p_r^{(m+1)} \psi_r(x_j; t_r^{(m+1)})}. \tag{9}$$

Таким образом, SEM-алгоритм представляет собой метод для оценивания неизвестных параметров компонент смеси без каких-либо дополнительных предположений об этих параметрах (например, нет необходимости в предположении о равенстве 0 параметра сдвига для каждой компоненты, то есть принудительного задания для модели вида строго масштабной смеси).

Как уже было отмечено, в работах, посвященных исследованию свойств SEM-алгоритма, предполагается выполнение ряда дополнительных условий. Так, в первых работах, посвященных данной тематике (см., например, статью [5]), устанавливались свойства лишь для двухкомпонентной смеси (при этом отмечалась невозможность обобщения приведенных доказательств на произвольное число компонент). В статье С. Ф. Нильсена [4] уже для произвольного числа компонент доказана сходимость SEM-алгоритма, установлены асимптотические свойства последовательности SEM-оценок (асимптотическая нормальность) при выполнении достаточно сложных для проверки на практике условий. Также предполагается строгая

положительность условной плотности (для сравнения см. формулу (9))

$$f_\theta(y_j \mid x_j) = \prod_{j=1}^{n} p_{y_j} \psi_{y_j}(x_j; t_{y_j}) \left(\sum_{i=1}^{k} p_i \psi_i(x_j; t_i) \right)^{-1}.$$

Данное условие существенным образом используется в статье [4] при доказательстве леммы об апериодичности и неразложимости SEM-цепи. Выше было отмечено, что в данной работе рассматриваются только такие x_j, которые принадлежат множеству $\mathfrak{D}$, однако плотность $f_\theta(y_j \mid x_j)$ может обращаться в нуль и при справедливости условия $p_{y_j} = 0$ для некоторого j. Этапы SEM-алгоритма непосредственно не запрещают весам обращаться в нуль, так что данное условие на практике может не выполняться. В следующем разделе будет сформулирована и доказана теорема о свойствах оценок SEM-алгоритма для произвольного числа компонент без введения дополнительных предположений о параметрах метода.

2. Эргодичность последовательности SEM-оценок

На каждом шаге работы алгоритма значения параметра $\theta^{(m)}$ можно представить в виде

$$\theta^{(m)} = f(\theta^{(m-1)}).$$

Однако эта функция $f(\cdot)$ зависит от величин $g_{ij}^{(m-1)}$, являющихся случайными величинами (из определения, см. формулу (8)). Поэтому последовательность $\{\theta^{(m)}\}$ является последовательностью *зависимых* случайных величин.

Введем ряд обозначений, которыми будем в дальнейшем пользоваться при работе с марковскими цепями. Обозначим вероятность перехода из состояния i в состояние j за k шагов через

$$p_{ij}(k) = \mathbb{P}(S_k = E_j \mid S_0 = E_i),$$

где S_l обозначает состояние цепи на l-ом шаге. В этих обозначениях элементы матрицы переходных вероятностей для однородной цепи (то есть матрицы, отражающей возможное изменение системы за один шаг) могут быть записаны как $p_{ij}(1)$, но для удобства будем обозначать их как p_{ij}. Через π_j обозначим стационарное распределение

$$\lim_{n\to\infty} p_{ij}(n) = \pi_j, \quad i = 1, \ldots, k,\ j = 1, \ldots, n.$$

Теперь можно установить свойства SEM-алгоритма и сформулировать результат в следующем виде.

Теорема 1. *Последовательность оценок $\{\theta^{(m)}\}$, получаемая SEM-алгоритмом в задаче разделения идентифицируемых смесей с произвольным конечным числом компонент, представляет собой конечную однородную апериодическую эргодическую марковскую цепь.*

Доказательство. Будем характеризовать каждое состояние SEM-цепи (то есть последовательности оценок $\left\{\theta^{(m)}\right\}$, получаемых на итерационных шагах) тем, в какой кластер $K_i^{(m)}$ на m-ом шаге попал элемент выборки x_j. Тогда каждое состояние можно однозначно характеризовать единственной неслучайной матрицей $Y(\eta)$ размера $k \times n$, η обозначает номер матрицы, элементами которой являются нули и единицы, причем $(Y(\eta))_{ij} = 1$ только в случае, когда элемент выборки x_j попал в кластер $K_i^{(m)}$ на m-ом шаге. Очевидно, что число таких матриц конечно и составляет k^n: можно рассмотреть соответствующий выборке набор длиной n, в котором на каждой позиции стоит номер кластера от 1 до k; общее число таких наборов, как известно, составляет k^n.

Тогда элементы данной матрицы определяют величины $g_{ij}^{(m)}$ на каждом шаге в соответствии с формулой (8): на $(m+1)$-м шаге можно выбрать такую неслучайную матрицу $Y(\eta)$, не зависящую от номера итерационного шага, что $(Y(\eta))_j \equiv \vec{y}_j^{(m+1)}$ (в силу конечного числа вариантов распределения элементов выборки по кластерам). Заметим, что фактически матрица $Y(\eta)$ представляет собой реализацию некоторой случайной матрицы $Y^{(m+1)}$, определяющей элементы $g_{ij}^{(m+1)}$ в формуле (8). Поэтому вероятность перехода из состояния S_m в состояние S_{m+1} однозначно определяется величиной $\mathbb{P}(S_{m+1} = Y(\eta) \mid S_m = Y(\nu))$ для некоторых номеров η и ν.

Итак, конечный набор матриц $Y(\eta)$ на каждом шаге однозначно (вместе с выборкой) определяет как значения элементов $g_{ij}^{(m)}$ (матрицу с элементами $g_{ij}^{(m)}$, $i = 1, \ldots, k$, $j = 1, \ldots, n$, соответствующую матрице $Y(\eta)$, будем обозначать через $G(\eta)$) (данная матрица представляет собой реализацию некоторой случайной матрицы $G^{(m+1)}$, элементы которой определяются формулой (8)), так и значения параметра $\theta^{(m)}$. Это означает как конечность числа матриц $G(\eta)$, так и конечность набора значений параметра $\theta^{(m)}$. Таким образом, можно выделить такое множество состояний, что вероятность перехода из текущего состояния в следующее, в силу определения SEM-алгоритма, не зависит от прошлого. Поэтому последовательность $\left\{\theta^{(m)}\right\}$ является марковской цепью.

Теперь рассмотрим вопрос определения элементов матрицы переходных вероятностей $p_{\nu\eta} = \mathbb{P}(S_{m+1} = Y(\eta) \mid S_m = Y(\nu))$, $\nu, \eta = 1, \ldots, k^n$. Введем обозначение

$$(G(\nu))^{Y(\eta)} \stackrel{\text{def}}{=} \prod_{i,j} (g_{ij}(\nu))^{y_{ij}(\eta)},$$

где

$$g_{ij}(\nu) = (G(\nu))_{ij}, \quad y_{ij}(\eta) = (Y(\eta))_{ij}.$$

Как уже было отмечено, состояние $Y(\eta)$ однозначно определяет матрицу $G(\eta)$. Из формулы (8) следует, что столбцы $(G(\nu))_j$ при разных j независимы (так как формула (8) справедлива и для реализаций случайных величин). Поэтому столбцы матрицы $(Y(\eta))_j$ независимы (так как по определению они являются реализациями случайных векторов с полиномиальным распределением с параметрами 1 и $(G(\nu))_j$). Это позволяет расписывать вероятность перехода как произведение вероятностей появления каждого столбца. Исходя из S-шага алгоритма, получаем, что

$$p_{\nu\eta} = (G(\nu))^{Y(\eta)}.$$

Из данного равенства видно, что элементы матрицы переходных вероятностей не зависят от номера итерации, а значит, последовательность оценок, которая строится SEM-алгоритмом, является однородной марковской цепью. Также из этого равенства следует (см. формулы (7) и (9)), что при условии разрешения существования пустых кластеров (то есть $g_{ij} = 0$ для всех $j = 1, \ldots, n$), если на каком-то шаге кластер i был объявлен пустым, то элементы вновь не могут на следующих шагах попадать в него (так как соответствующий элемент g_{ij} равен нулю и с ненулевой вероятностью возможен переход только в состояния с $y_{ij} = 0$, то есть элемент выборки x_j кластеру i не принадлежит).

Каждая вероятность $p_{\nu\eta}$ представляет собой произведение n множителей g_{ij}, отвечающих распределению элементов выборки по кластерам. Из определения очевидно, что $p_{\nu\eta} \geqslant 0$ для всех возможных ν и η (в том смысле, что допускаются и нулевые вероятности перехода из одного состояние в другое; как было отмечено выше, если цепь перешла в состояние, в котором некоторый кластер считается пустым, перейти в состояния, в которых этот кластер содержит какие-то элементы выборки, невозможно).

Правило суммирования. На l-ом шаге группируем по k штук слагаемых с одинаковыми $n - l$ членами. Затем пользуемся соотношением (6) – и таким образом получаем итоговую сумму, равную единице.

Итак, $\sum_\eta p_{\nu\eta} = 1$. Сказанное означает, что SEM-цепь $\left\{\theta^{(m)}\right\}$ является конечной однородной марковской цепью со стохастической матрицей перехода. При этом, если вероятность перехода из состояния i в состояние j не равна 0, то возможно перейти в любое состояние в точности за 1 шаг с ненулевой вероятностью. Поэтому SEM-цепь апериодична.

Пусть сначала выполнено условие $g_{ij} > 0$ для всех возможных номеров i и j (то есть запрещается ситуация, когда какой-то кластер признается пустым). Тогда матрица переходных вероятностей получается апериодической, неразложимой (все состояния цепи являются существенными, сообщающимися, так как с положительной вероятностью можно попасть из одного произвольного состояния в другое), что в силу ее конечности согласно результатам книги [6] означает эргодичность цепи. В этом случае разность $|p_{ij}(n) - \pi_j|$ имеет экспоненциальную скорость сходимости к нулю (то есть получаем экспоненциальную скорость сходимости к эргодическому распределению).

Пусть теперь выполнено условие $g_{ij} \geqslant 0$ для всех возможных номеров i и j (то есть разрешается ситуация, когда какой-то кластер признается пустым). Обозначим через $\{E_i\}$ множество состояний SEM-цепи. Выделим в отдельное состояние ситуацию, когда заполнен единственный кластер, а остальные являются пустыми, то есть в этой ситуации $(Y(\eta))_{ij} = 1$ только для одного значения i и сразу для всех j. Этим реализуется понятное с практической точки зрения правило, означающее, что значения параметров не зависят от того, какой именно единственный кластер заполнен, ибо изначально наша нумерация условна и допускает переобозначение. Обозначим это состояние E_1. Соответствующую вероятность перехода в это состояние положим равной сумме вероятностей переходов в каждое из подобных состояний (соответствующих матрицам $Y(\eta)$ для таких значений η, что единственная строка в них полностью заполнена единицами, а остальные элементы – нули). Таким образом, сохраним первоначальную взаимосвязь состояний и

стохастичность матрицы перехода.

Очевидно (в силу апериодичности и достижимости всех состояний), что существует $t_0 > 0$ такое, что для любого $i \geqslant 1$ верно, что вероятность $p_{i1}(t_0) > 0$, но условие $p_{1i}(t) > 0$ для любого $t_0 > 0$ влечет условие $i = 1$ (как уже говорилось, кластер не может содержать элементы, если однажды он оказался пустым). Иначе $p_{1i}(t) \equiv 0$ для любого $t_0 > 0$. По определению это означает, что все состояния, кроме E_1, являются несущественными. При этом понятно, что E_1 – существенное. То есть множество состояний разбивается на два класса: множество всех несущественных состояний S^0 (состоит из всех E_i, $i > 1$) и множество существенных состояний S^1 (состоит из E_1). В данном случаем получаем, что SEM-цепь является разложимой. Поэтому воспользоваться результатами для случая $g_{ij} > 0$ уже нельзя.

Запишем матрицу переходных вероятностей в виде

$$\begin{array}{c} \\ S^0 \\ S^1 \end{array} \begin{array}{c} \begin{array}{cc} S^0 & S^1 \end{array} \\ \begin{pmatrix} ///// & ///// \\ 0 & L \end{pmatrix} \end{array}.$$

Здесь матрица $L \equiv 1$ и, очевидно, является стохастической, неразложимой и апериодической, что в случае конечности цепи означает ее эргодичность. Понятно, что S^1 не содержит подклассов. Такая структура матрицы переходных вероятностей является необходимым и достаточным условием того, чтобы была справедлива следующая лемма [6].

Лемма 1. *Если* $E_i \in S^0$, *то*

$$\lim_{t\to\infty} p_{ij}(t) = \begin{cases} 0, & \text{если } E_j \in S^0, \\ \pi_j, & \text{если } E_j \in S^1. \end{cases}$$

Воспользовавшись результатами этой леммы, можно говорить об эргодичности разложимой цепи. В силу эргодичности матрицы L,

$$\lim_{n\to\infty} p_{i1}(n) = \pi_1,$$
$$\sum_n \pi_1 = 1,\ \pi_1 = (\mathbb{E}\tau)^{-1},$$

где $\mathbb{P}(\tau^{(1)} = n) = f_1(n)$, $\tau^{(1)}$ – время возвращения в E_1, а

$$f_1(n) = \mathbb{P}(S_n = E_1 \mid S_{n-1} \neq E_1, ..., S_1 \neq E_1 \mid S_0 = E_1)$$

представляет собой вероятность того, что если система вышла из состояния E_1, то она впервые вернется в него за n шагов. Но если цепь *уже* находится в состоянии E_1, то переход в E_1 за один шаг осуществляется с вероятностью 1. Поэтому $\mathbb{P}(\tau^{(1)} = 1) = 1$ и $\mathbb{E}\tau^{(1)} = 1$. Откуда следует, что $\pi_1 = 1$. □

Заключение

Итак, теорема 1 описывает сходимость последовательности оценок параметров произвольной конечной смеси вероятностных распределений, генерируемой SEM-

алгоритмом, в терминах эргодичности. Из этой теоремы вытекает, что в случае допустимости существования пустых кластеров SEM-цепь становится стационарной только при попадании в поглощающее состояние E_1. При этом можно говорить о поточечной сходимости последовательности $\{\theta^{(m)}\}$: так как начиная с момента попадания в состояние E_1 справедливо тождество $\theta^{(n)} = \theta^{(m)}$ сразу для всех последующих номеров итерационных шагов n и m. Следовательно, можно говорить о сходимости с любой наперед заданной точностью. При этом скорость сходимости SEM-алгоритма в случае допустимости существования пустых кластеров определяется временем попадания в состояние E_1. В качестве оценки параметров при допустимости пустых кластеров при достижении стационарного режима берутся оценки выборки по соответствующим конкретному типу смеси формулам. Так можно оценить компоненту, вносящую наиболее значительный вклад в портрет волатильности. При этом запрет существования пустых кластеров позволяет подгонять к данным модель строго k-компонентной смеси, что оказывается удобным при статистической проверке числа компонент в смеси (подробнее см. статьи [7, 8]). Такой подход позволяет точнее оценивать число компонент в смеси, а значит, делает возможным выбор модели, адекватно описывающей распределение исходной выборки. В заключение отметим, что для некоторых важных типов смесей вероятностных распределений, например, для конечных смесей нормальных законов, проводились эмпирические испытания, иллюстрирующие теоретические результаты (более подробно на примере финансовых индексов и турбулентной плазмы об этом в статьях [9, 10]).

Список литературы

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

[2] Diebolt J., Ip E. H. Stochastic EM: method and application // W. R. Gilks, S. Richardson, D. J. Spiegelhalter (Eds.) Markov Chain Monte Carlo in Practice. – London: Chapman and Hall, 1996.

[3] Ip E. H. A Stochastic EM Estimator in the Presence of Missing Data. – Theory and Practice. PhD Dissertation, Stanford University, 1994.

[4] Nielsen S. F. Stochastic EM algorithm: Estimation and asymptotic results // Bernoulli, 2000. № 6. P. 457–489.

[5] G. Celeux, J. Diebolt. Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions // Communications in statistics. Stochastic models 1993. Vol. 9. P. 599–613.

[6] Боровков А. А. Теория вероятностей. Изд. 4-е – М.: Едиториал УРСС, 2003.

[7] Бенинг В. Е., Горшенин А. К., Королев В. Ю. Асимптотически оптимальный критерий проверки гипотез о числе компонент смеси вероятностных распределений // Информатика и ее применения, 2011. Т. 5. Вып. 3. С. 4–16.

[8] Горшенин А. К. Проверка статистических гипотез в модели расщепления компоненты // Вестник Московского Университета, 2011. Серия 15. Вычислительная математика и кибернетика. № 4. С. 26–32.

[9] 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.

[10] 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.