А. К. Г о р ш е н и н (Москва, МГУ)**. Проверка гипотез о числе компонент смеси вероятностных распределений.**

В важных прикладных задачах, использующих математическую модель конечных смесей вероятностных распределений (анализ финансовых рынков, турбулентной плазмы, см. книгу [1]), необходима корректная интерпретация полученных результатов. Поэтому определяющее значение имеет выбор числа компонент в модели, подгоняемой к реальным данным. Пусть k — некоторое известное натуральное число. Требуется проверить гипотезу H_0: $K = k$ против альтернативы H_1: $K = k + 1$, где K обозначает «истинное» число компонент в смеси. Для удобства асимптотического анализа предлагаемых критериев сведем задачу проверки гипотез о значении *дискретного* параметра K к задаче проверки гипотез о значении *непрерывного* параметра: рассматривается простая гипотеза вида H_0: $\theta = 0$ против сложной альтернативы вида H_1: $\theta > 0$. Для нахождения асимптотически наиболее мощного критерия используется подход, с помощью которого можно заменить сложную альтернативу последовательностью простых вида $H_{n,1}$: $\theta = t/\sqrt{n}$, $0 < t \leqslant C$, $C > 0$ (см., например, [2]). Отметим, что всюду мы предполагаем, что исходная смесь является идентифицируемой.

В большинстве популярных алгоритмов для статистической декомпозиции смесей (EM-алгоритм, его всевозможные модификации, сеточные методы) используется заранее заданное число компонент. В такой ситуации необходимо убедиться в значимости компоненты с малым весом или отбросить ее без потери информативности модели. Предположим, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ имеет плотность, представимую для некоторого $\theta \in [0, 1]$ в виде (все $p_i \geqslant 0$, $i = 1, 2, \ldots, k$)

$$p(x, \theta) = (1 - \theta) \sum_{i=1}^{k} p_i \psi_i(x) + \theta \psi_{k+1}(x) = (1 - \theta) f(x) + \theta g(x), \quad \sum_{i=1}^{k} p_i = 1. \qquad (1)$$

Тогда справедлива следующая теорема.

Теорема 1. *Пусть моментные характеристики* $\Psi_s = \mathbf{E}_0 (g(X_1) f^{-1}(X_1))^s$, $s = 2, 3, 4$, *для функций* $f(x)$ *и* $g(x)$ *из соотношения* (1) *конечны. Тогда для проверки гипотез о числе компонент в модели добавления компоненты* (1) *для идентифицируемой смеси законов вероятностных распределений может быть использован критерий, основанный на статистике* $L_n^{(1)} = n^{-1/2} \sum_{i=1}^{n} (g(X_i) f^{-1}(X_i) - 1)$ *и обладающий следующими свойствами.*

1) *При справедливости нулевой гипотезы статистика* $L_n^{(1)}$ *имеет нормальное распределение с параметрами* 0 *и* $\Psi_2 - 1$: $\mathfrak{L}(L_n^{(1)} \mid H_0) \to N(0, \Psi_2 - 1)$.

2) *При справедливости альтернативы статистика* $L_n^{(1)}$ *имеет нормальное распределение с параметрами* $t(\Psi_2 - 1)$ *и* $\Psi_2 - 1$: $\mathfrak{L}(L_n^{(1)} \mid H_{n,1}) \to N(t(\Psi_2 - 1), \Psi_2 - 1)$.

3) *Данный критерий является асимптотически наиболее мощным критерием для заданного уровня* $\alpha \in (0, 1)$ *с предельной мощностью вида* $\beta^*(t) = \Phi(t\sqrt{\Psi_2 - 1} - u_\alpha)$.

4) *Потеря мощности этого критерия равна*

$$r(t) = \frac{t^3}{8\sqrt{\Psi_2 - 1}} \varphi(u_\alpha - t\sqrt{\Psi_2 - 1}) \left(\Psi_4 + 2\Psi_3 - \Psi_2^2 - \Psi_2 - \frac{(\Psi_3 - 1)^2}{\Psi_2 - 1} - 1 \right).$$

5) *Асимптотический дефект этого критерия равен*

$$d = \frac{t^2}{4(\Psi_2 - 1)} \left(\Psi_4 + 2\Psi_3 - \Psi_2^2 - \Psi_2 - \frac{(\Psi_3 - 1)^2}{\Psi_2 - 1} - 1 \right).$$

Здесь $\Phi(u_\alpha) = 1 - \alpha$, символ $\Phi(\cdot)$ обозначает функцию распределения стандартного нормального закона.

На практике часто возникает ситуация, связанная с еще одной особенностью наиболее часто используемых методов статистического разделения смесей. Дело в том, что алгоритмы данного типа в ряде ситуаций ошибочно объединяют близкие по параметрам компоненты в одну общую, хотя на самом деле это не так (см., например, статью [3]). Возникает необходимость проверить значимость нескольких компонент с близкими параметрами или объединить их без потери информативности модели. Предположим, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ имеет плотность, представимую для некоторого $\theta \in [0, 1]$ в виде (все $p_i \geqslant 0$, $i = 1, 2, \ldots, k$)

$$p(x, \theta) = \sum_{i=1}^{k} p_i \psi_i(x) + \theta(\psi(x) - \psi_k(x)) = f(x) + \theta g(x), \quad \sum_{i=1}^{k} p_i = 1. \tag{2}$$

В этом случае справедлива следующая теорема.

Теорема 2. *Пусть моментные характеристики* $\Psi_s = \mathbf{E}_0(g(X_1)f^{-1}(X_1))^s$, $s = 2, 3, 4$, *для функций* $f(x)$ *и* $g(x)$ *из соотношения* (2) *конечны. Тогда для проверки гипотез о числе компонент в модели расщепления компоненты* (2) *идентифицируемой смеси законов вероятностных распределений может быть использован критерий, основанный на статистике* $L_n^{(1)} = n^{-1/2} \sum_{i=1}^{n} [g(X_i)/f(X_i)]$ *и обладающий следующими свойствами.*

1. *Статистика* $L_n^{(1)}$ *асимптотически нормальна,* $\mathfrak{L}(L_n^{(1)} \mid H_0) \to N(0, \Psi_2)$, $\mathfrak{L}(L_n^{(1)} \mid H_{n,1}) \to N(t\Psi_2, \Psi_2)$.

2. *Данный критерий является асимптотически наиболее мощным критерием для заданного уровня* $\alpha \in (0, 1)$ *с предельной мощностью вида* $\beta^*(t) = \Phi(t\sqrt{\Psi_2} - u_\alpha)$.

3. *Потеря мощности для этого критерия составляет*

$$r(t) = \frac{t^3}{8\sqrt{\Psi_2}} \varphi(u_\alpha - t\sqrt{\Psi_2}) \left(\Psi_4 - \Psi_2^2 - \frac{\Psi_3^2}{\Psi_2} \right).$$

4. *Асимптотический дефект для этого критерия равен*

$$d = \frac{t^2}{4\Psi_2} \left(\Psi_4 - \Psi_2^2 - \frac{\Psi_3^2}{\Psi_2} \right).$$

СПИСОК ЛИТЕРАТУРЫ

1. *Королев В. Ю.* Вероятностно-статистический анализ хаотических процессов с помощью смешанных гауссовских моделей. Декомпозиция волатильности финансовых индексов и турбулентной плазмы. М.: ИПИ РАН, 2007, 363 с.
2. *Bening V. E.* Asymptotic Theory Of Testing Statistical Hypothesis: Efficient Statistics, Optimality, Power Loss and Deficiency. Untrecht: VSP, 2000, 277 p.
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 во латильности финансовых временных рядов. – Информатика и ее применения, 2008, т. 2, в. 4, с. 12–47.