УДК 519. 22

Р. Є. Майборода (Київ. ун-т)

ОЦІНКА РОЗПОДІЛІВ КОМПОНЕНТ СУМІШЕЙ З КОНЦЕНТРАЦІЯМИ, ЩО ЗМІНЮЮТЬСЯ

For the data choosed from a mixture with varying concentrations, nonparametric estimations for distributions of components and Spearman correlation coefficient are constructed. Consistency of the correlation coefficient and efficiency of the distribution estimations are proved.

Для даних, що є вибіркою з суміші кількох компонент, концентрація яких змінюється протягом спостережень, побудовано непараметричні оцінки розподілів компонент та ранговий коефіцієнт кореляції. Доведено спроможність рангового коефіцієнта та ефективність оцінок розподілів.

Задачі статистичного аналізу вибірок із сумішей кількох компонент з різними розподілами дуже часто виникають при обробці біологічних, психологічних та соціально-економічних даних. Параметричним методам розщеплення сумішей присвячено роботи [1, 2].

У даній роботі встановлена непараметрична оцінка розподілів компонент сумішей за умови, що концентрації компонент змінюються з часом спостереження і відомі. Про оцінку концентрацій див. [3, 4].

Нехай $(\Omega, \mathfrak{T}, \Pr)$ — основний імовірнісний простір, $(\Delta, \mathfrak{A})$ — вимірний простір спостережень, $\xi_j^N : \Omega \to \Delta$ — вимірні відображення (випадкові елементи, що набувають значень в Δ), $j = 1 \div N$, $N \in \mathbb{N}$.

З кожним спостереженням ми пов'язуємо число $t_j^N \in [0, 1]$, яке називаємо часом, коли проводилось спостереження. Набір $Q_N = \{ t_j^N, j = 1 \div N \}$ будемо називати сіткою часу. Вважаємо, що $0 = t_1^N < t_2^N < \ldots < t_N^N = 1$.

Вважається, що при фіксованому N випадкові елементи ξ_j^N, $j = 1 \div N$, незалежні в сукупності і

$$\Pr \{ \xi_j^N \in A \} = \mu(A, t_j^N) = \sum_{k=1}^{M} \omega_k(t_j^N) H_k(A), \tag{1}$$

де M — кількість компонент у суміші; $\omega_k(t)$ — концентрація k-ї компоненти суміші; $H_k(\cdot)$ — розподіл k-ї компоненти суміші.

Нехай $a : [0, 1] \to \mathbb{R}$ — деяка функція. Емпіричною мірою (е.м.) [5] з вагою a, побудованою за спостереженнями Ξ, назвемо

$$\hat{\mu}_N(A, a) = \frac{1}{N} \sum_{j=1}^{N} a(t_N^j) \chi \{ \xi_j^N \in A \}, \tag{2}$$

де $\chi\{U\}$ — індикатор події U.

Подібно тому, як у випадку однорідної вибірки для оцінки її розподілу звичайно використовують однорідну емпіричну міру (див. [6, 7]), для оцінки розподілу компонент в (1) можна використати (2) з ваговою функцією, вибраною в залежності від функції концентрації ω_k. При цьому природно вимагати, щоб оцінка була незміщеною, тобто $\mathbf{E}\, \hat{\mu}_N(A, a_l) = H_l(A)$ для будь-яких A. Позначимо

$$\langle a, b \rangle_N = \frac{1}{N} \sum_{j=1}^{N} a(t_j^N)\, b(t_j^N).$$

Тоді з (1) маємо

ISSN 0041-6053. Укр. мат. журн., 1996, т. 48, № 4

$$\mathbf{E}\,\hat{\mu}_N(A, a_l) = \sum_{k=1}^{M} H_k(A)\langle a_l, \omega_k\rangle_N.$$

Очевидно, що $\mathbf{E}\,\hat{\mu}_N(A, a_l)_N = H_l(A)$ при будь-яких A і $H_k(A)$ в тому і тільки в тому випадку, коли

$$\langle a_l, \omega_k\rangle_N = \chi\{l = k\}. \tag{3}$$

Таким чином, (3) є необхідною та достатньою умовою незміщеності $\hat{H}_l(A) = \hat{\mu}_N(A, a_l)$ як оцінки для розподілу k-ї компоненти суміші.

Серед усіх функцій a_l, що задовольняють (3), виберемо таку, що мінімізує гарантований ризик при квадратичній функції витрат. Точніше, будемо вважати, що $\mathbf{E}\,(\hat{H}_l(A) - H_l(A))^2$ характеризує середні витрати, що виникають при заміні $H_l(A)$ її оцінкою $\hat{H}_l(A)$. Тоді

$$J(a_l) = R(\hat{H}_l) = \sup_{H_l,\, A} \mathbf{E}(\hat{H}_l(A) - H_l(A))^2 \tag{4}$$

є гарантованим ризиком. (Супремум тут беремо по всіх можливих розподілах H_l та всіх $A \in \mathfrak{A}$). Враховуючи (3), маємо

$$J(a_l) = \sup_{H_l,\, A}\left(\frac{1}{N^2}\sum_{j=1}^{N}\left(a(t_j^N)\right)^2\left(\Pr\{\xi_j^N \in A\} - \left(\Pr\{\xi_j^N \in A\}\right)^2\right)\right) =$$

$$= \frac{1}{4N}\langle a_l, a_l\rangle_N,$$

оскільки

$$\sup_{H_l,\, A}\left(\Pr\{\xi_j^N \in A\} - \left(\Pr\{\xi_j^N \in A\}\right)^2\right) \le \frac{1}{4}$$

і значення $1/4$ досягається при $H_l(A) = 1/2$.

Таким чином, оптимальна функція a_l мінімізує $\langle a_l, a_l\rangle_N$ при умові (3). Розв'язуючи цю оптимізаційну задачу методом множників Лагранжа, отримуємо

$$a_l(t) = \frac{1}{\det \Gamma_N}\sum_{k=1}^{M}(-1)^{l+k}\gamma_{lk}^N\omega_k(t), \tag{5}$$

де $\Gamma_N = (\langle\omega_k, \omega_l\rangle_N)_{k,l=1}^M$ — матриця Грама системи функцій (векторів) ω_k $k = 1 \div M$, у скалярному добутку $\langle\cdot,\cdot\rangle_N$, γ_{lk}^N — (l, k)-й мінор матриці Γ_N.

Зрозуміло, що (5) має сенс лише тоді, коли $\det\Gamma_N \neq 0$. Умова $\det\Gamma_N = 0$ рівносильна лінійній залежності ω_k, $k = 1 \div M$, як функцій на множині $\{t_j^N,\ j = 1 \div N\}$. Лінійна залежність функцій концентрації робить неоднозначним розклад (1), отже, в цьому випадку оцінка розподілів компонент суміші не має сенсу (вони теж не визначені однозначно). Наприклад, якщо вважати концентрації незалежними від часу спостережень, отримуємо класичну задачу розділення суміші M компонент, яка в непараметричній постановці є некоректною.

Покажемо, що відносно ризику (4) наша оцінка є ефективною на класі всіх незміщених оцінок.

Теорема 1. *Нехай спостереження* $\Xi_N = \{\xi_1^N, \ldots, \xi_N^N\}$ *визначені* (1), $\omega_k(t)$ *вважаються відомими,* H_k *— невідомими. Якщо* $\tilde{H}_l^N : \mathfrak{A} \times \Delta^N \to [0, 1]$

— вимірна функція, така, що для всіх $A \in \mathfrak{A}$ *та всіх можливих розподілів* H_k, $\mathbf{E}\,\tilde{H}_l^N(A, \Xi_N) = H_l(A)$, *то*

$$R(\tilde{H}_l(\cdot, \Xi_N)) \geq J(a_l),$$

де a_l *визначено* (5).

Доведення. Виберемо довільні $x_1, x_2 \in \Delta, x_1 \neq x_2$. Нехай H_k — розподіл вигляду

$$H_k(A) = \begin{cases} 0, & \text{якщо} \quad \{x_1, x_2\} \cap A = \varnothing; \\ p_k, & \text{якщо} \quad x_1 \in A, x_2 \notin A; \\ 1 - p_k, & \text{якщо} \quad x_2 \in A, x_1 \notin A; \\ 1, & \text{якщо} \quad \{x_1, x_2\} \in A. \end{cases} \tag{6}$$

Якщо обмежитися розглядом лише сумішей (1) з розподілами компонент виду (6), то задача оцінки $H_k(A)$ зведеться до оцінювання вектора $\vec{p} = (p_1, \ldots, p_M)$. Це задача параметричного оцінювання.

Підрахуємо інформаційну матрицю Фішера I для стохастичного експерименту, який полягає в тому, що за спостереженням ξ_j^N оцінюється $H_k(\{x_1\}) = p_k$, $k = 1 \div M$, тоді як їх справжніми значеннями є $p_1^0 = p_2^0 = \ldots = p_M^0 = 1/2$ [8, с. 91]. Для цього введемо міру

$$\nu(A) = \begin{cases} 0, & \text{якщо} \quad \{x_1, x_2\} \cap A = \varnothing; \\ 1, & \text{якщо} \quad x_1 \in A, x_2 \notin A \text{ або } x_2 \in A, x_1 \notin A; \\ 2, & \text{якщо} \quad \{x_1, x_2\} \in A. \end{cases}$$

Зрозуміло, що H_k абсолютно неперервні відносно ν при будь-яких $\vec{p}$ та k і

$$h_k(x, \vec{p}) = \frac{dH_k}{d\nu} = \begin{cases} p_k, & \text{якщо} \quad x = x_1; \\ 1 - p_k, & \text{якщо} \quad x = x_2; \\ 0, & \text{якщо} \quad x \notin \{x_1, x_2\}. \end{cases}$$

Елементи інформаційної матриці $I^j = \left(I_{kl}^j\right)_{k,l=1}^M$ для спостереження ξ_j^N можна підрахувати за формулою

$$I_{kl}^j = \int_\Delta \frac{\partial h(x, \vec{p})}{\partial p_k} \frac{\partial h(x, \vec{p})}{\partial p_l} \frac{\nu(dx)}{h(x, \vec{p})} \Bigg|_{\vec{p} = \vec{p}^0},$$

де

$$h(x, \vec{p}) = \frac{d\mu(\cdot, t_j^N)}{d\nu} = \sum_{i=1}^M \omega_i(t_j^N)\, h_i(x, \vec{p}).$$

Легко бачити, що

$$I_{kl}^j = \frac{\omega_k(t_j^N)\,\omega_l(t_j^N)}{\sum_{i=1}^M \omega_i(t_j^N) p_i^0} + \frac{\omega_k(t_j^N)\,\omega_l(t_j^N)}{1 - \sum_{i=1}^M \omega_i(t_j^N) p_i^0} = 4\,\omega_k(t_j^N)\,\omega_l(t_j^N).$$

Оскільки спостереження ξ_j^N назалежні при фіксованому N, то $I = \sum_{j=1}^N I^j = 4N\Gamma_N$. Згідно з нерівністю Крамера–Рао [8, с. 104], враховуючи незміщеність оцінки $\tilde{H}_l^N(\cdot, \Xi_N)$, маємо

$$\mathbf{E}\,(\tilde{H}_l^N(\{x_1\}, \Xi_N) - H_l(\{x_1\}))^2 \;\geq\; e_l^T I^{-1} e_l \;=\; \frac{e_l^T \Gamma_N^{-1} e_l}{4N},$$

де $e_l = (e_l^1, \ldots, e_l^M)^T$, $e_l^k = \chi\{l = k\}$.

Оскільки супремум в означенні $R(\tilde{H}_l^N(\cdot, \Xi_N))$ береться по класу, що включає в себе розподіли H_k і множину $A = \{x_1\}$, то

$$R(\tilde{H}_l^N(\cdot, \Xi_N)) \;\geq\; e_l^T I^{-1} e_l \;=\; \frac{e_l^T \Gamma_N^{-1} e_l}{4N}. \tag{7}$$

Обчислимо тепер $J(a_l) = \dfrac{1}{4N}\langle a_l, a_l\rangle_N$ для a_l, визначеного (5). Зауважимо, що $a_l(t) = e_l^T \Gamma_N^{-1} \bar{\omega}(t)$, де $\bar{\omega}(t) = (\omega_1(t), \ldots, \omega_M(t))^T$. Таким чином,

$$\langle a_l, a_l\rangle = \left\langle e_l^T \Gamma_N^{-1} \bar{\omega}(\cdot),\, e_l^T \Gamma_N^{-1} \bar{\omega}(\cdot)\right\rangle_N \;=\; e_l^T \Gamma_N^{-1} \Gamma_N \Gamma_N^{-1} e_l \;=\; e_l^T \Gamma_N^{-1} e_l,$$

звідки, враховуючи (7), маємо $R(\tilde{H}_l^N(\cdot, \Xi_N)) \geq J(a_l)$, що й треба було довести.

Оцінки розподілів компонент, альтернативні ефективним, можна будувати за допомогою методу емпіричного максимуму правдоподібності (ЕМП) [9 – 11]. ЕМП-оцінювання приводить до зважених емпіричних мір з позитивними ваговими функціями. Ці оцінки не є ні (асимптотично) незміщеними, ні навіть спроможними.

Як правило, спеціалістів у предметних областях цікавлять деякі функціонали від H_k. Маючи оцінку $\hat{H}_k$ для H_k, можна будувати і відповідні оцінки для функціоналів. Прикладом може бути ранговий коефіцієнт кореляції Спірмена, який ми тут і розглянемо. Опишемо його модифікацію, орієнтовану на вибірку з суміші зі змінними концентраціями. Нехай $\Delta = \mathbb{R}^2$ і, отже, $\xi_j^N = (\xi_j^N(1), \xi_j^N(2))$ — двовимірний вектор. Розглянемо коефіцієнт, який характеризує кореляцію між першою та другою координатами цього вектора, за умови, що об'єкт спостереження належить k-му класу (тобто цей коефіцієнт характеризує кореляцію для H_k). Введемо спочатку поняття рангу j-го спостереження по l-й координаті для k-ї компоненти:

$$r_j^{lk} \;=\; N\hat{\mu}_N\left(\left\{(x_1, x_2) \in \mathbb{R}^2 \mid\; x_l \leq \xi_j^N(l)\right\}, a_k(\cdot)\right) =$$

$$= \sum_{i=1}^{N} a_k(t_i^N)\,\chi\{\xi_i^N(l) \leq \xi_j^N(l)\}.$$

Узагальнений коефіцієнт кореляції між першою та другою координатами k-ї компоненти визначається як

$$\rho_N^k \;=\; 1 - \frac{6}{N^3 - N}\sum_{i=1}^{N} a_k(t_i^N)(r_i^{1k} - r_i^{2k})^2.$$

Поклавши $a_k \equiv 1$, отримуємо звичайний коефіцієнт Спірмена. У термінах неоднорідних емпіричних мір маємо

$$\rho_N^k \;=\; \rho^k(\hat{F}_N^k) \;=$$

$$= 1 - \frac{6N^2}{N^2 - 1}\int_{\mathbb{R}^2} (\hat{F}_N(x_1, +\infty) - \hat{F}_N(+\infty, x_2))^2\, \hat{F}_N(dx_1, dx_2), \tag{8}$$

де $\hat{F}_N(x_1, x_2) = \hat{\mu}_N((-\infty, \bar{x}], a_k(\cdot))$, $\bar{x} = (x_1, x_2)$. Якщо

1) $\{\omega_l, l = 1 \div M\}$ — лінійно незалежні в $L_2[0, 1]$ функції з обмеженими варіаціями і

2) $t_j^N = j/N$,

то за теоремою 1 з [12] при $N \to \infty$ $\hat{F}_N(x_1, x_2)$ рівномірно по $x_1, x_2 \in \mathbb{R}$ збігається м.н. до $H_k((-\infty, \bar{x}])$. Легко переконатись, що неперервні функції розподілу у $\mathbb{R}^2$ є точками неперервності функціоналу $\rho^k(\cdot)$, визначеного (8), відносно рівномірної метрики. Отже справедлива наступна теорема.

Теорема 2. *Якщо виконані умови* 1, 2, *розподіли* H_k, $k = 1 \div M$, *неперервні, то при* $N \to \infty$

$$\rho_N^k \to \rho_\infty^k(H_k) = 1 - 6 \int (H_k(x_1, \infty) - H_k(\infty, x_2))^2 H_k(dx_1, dx_2).$$

1. *Teicher H.* Identifiability of mixtures // Ann. Math. Statist. – 1961. – **32**, № 1 – P. 244–248.
2. *Айвазян С. А. и др.* Прикладная статистика: Классификация и снижение размерности. – М.: Финансы и статистика, 1989. – 607 с.
3. *Майборода Р. Є.* Проекційні оцінки концентрацій сумішей, що змінюються // Теория вероятностей и мат. статистика. – 1992. – Вип. 44. – С. 70–76.
4. *Maiboroda R. E.* Estimators for parameters of timedependent mixture concentrations // Допов. НАН України. – 1993. – № 4. – P. 34–37.
5. *Stone C. J.* Consistent nonparametric regression // Ann. Statist. – 1977. – **5**. – P. 595–645.
6. *Биллингсли П.* Сходимость вероятностных мер. – М.: Наука, 1977. – 352 с.
7. *Dudly R. M.* Central limit theorems for empirical measures // Ann. Probab. – 1978. – **6**, № 6. – P. 899–929.
8. *Ибрагимов И. А., Хасьминский Р. З.* Асимптотическая теория оценивания. – М.: Наука, 1979. – 528 с.
9. *Vardi Y.* Empirical distributions in selection bias models // Ann. Statist. – 1985. – **13**, № 1. – P. 178–203.
10. *Owen A.* Empirical likelihood ratio confidence intervals for single functional // Biometrika. – 1988. – **75**. – P. 237–249.
11. *Gill R. D., Van der Vaart A. W.* Non- and semi-parametric maximum likelihood estimators and the von Mises method II // Scand. J. Statist. – 1993. – **20**. – P. 271–288.
12. *Майборода Р. Е.* Об оценивании параметров переменных смесей // Теория вероятностей и мат. статистика. – 1991. – Вып. 44. – С. 87–92.
13. *Боровков А. А.* Математическая статистика. – М.: Наука, 1984. – 472 с.

Одержано 06.12.94