УДК 519.21

В. Л. Гирко, Т. В. Павленко

G-оценка квадратичной дискриминантной функции

Настоящая работа посвящена построению и изучению свойств G-оценки квадратичной дискриминантной функции в случае двух многомерных нормальных генеральных совокупностей. Вопросы применения методов общего статистического анализа (G-анализа) к построению оценок некоторых статистик многомерного статистического анализа рассматривались в работах [1—4].

Рассмотрим задачу классификации случайного m-мерного вектора x, т. е. задачу отнесения его к той или иной совокупности, соответствующей одному из нормальных распределений $N(a_1, R_1)$ и $N(a_2, R_2)$, a_1, a_2— векторы средних значений, R_1, R_2 — ковариационные матрицы. Будем предполагать, что априорные вероятности наблюдения над совокупностями $N(a_1, R_1)$ и $N(a_2, R_2)$ равны. Цены неправильной классификации также равны. В этом случае квадратичная дискриминантная функция имеет следую-

щий вид:

$$F(x) = (x - a_1)' R_1^{-1} (x - a_1) - (x - a_2)' R_2^{-1} (x - a_2) + \ln \det R_2 (\det R_1)^{-1}.$$

Пусть имеются обучающие выборки $x_1^{(1)}, \ldots, x_{n_1}^{(1)} \in N(a_1, R_1)$, $x_1^{(2)}, \ldots, x_{n_2}^{(2)} \in N(a_2, R_2)$. Ранее [5] в качестве оценки $F(x)$, как правило, рассматривалось выражение

$$\hat{F}(x) = (x - \hat{a}_1)' \hat{R}_1^{-1} (x - a_1) - (x - \hat{a}_2)' \hat{R}_2^{-1} (x - \hat{a}_2) + \ln \frac{\det \hat{R}_2}{\det \hat{R}_1},$$

где

$$\hat{a}_1 = n_1^{-1} \sum_{i=1}^{n_1} x_i^{(1)}, \quad \hat{a}_2 = n_2^{-1} \sum_{i=1}^{n_2} x_i^{(2)}, \quad \hat{R}_1 = (n_1 - 1)^{-1} \sum_{i=1}^{n_1} (x_i^{(1)} - \hat{a}_1)(x_i^{(1)} - \hat{a}_1)',$$

$$\hat{R}_2 = (n_2 - 1)^{-1} \sum_{i=1}^{n_2} (x_i^{(2)} - \hat{a}_2)(x_i^{(2)} - \hat{a}_2)'.$$

Предполагалось, что выполняется условие $\lim\limits_{n_i \to \infty} m n_i^{-1} = 0$. Очевидно, что при больших размерностях наблюдаемого вектора необходимо значительное количество наблюдений, что существенно осложняет использование статистики $\hat{F}(x)$ на практике. В связи с этим более полезным оказался подход, основанный на асимптотике А. Н. Колмогорова [6]

$$\lim_{n_i \to \infty} m n_i^{-1} = c_i, \quad 0 < c_i < \infty, \quad c_i = \text{const}. \tag{1}$$

При условии (1) стандартные оценки неприемлемы. Применяя методы общего статистического анализа, можно построить оценки, использование которых позволит значительно сократить количество наблюдений при решении практических задач.

Оценки, которые являются состоятельными при выполнении условия А. Н. Колмогорова, будем называть G-оценками.

Покажем, что при некоторых условиях в качестве G-оценки квадратичной дискриминантной функции можно взять выражение

$$G_F(x) = A_1(x) + A_2(x) - B_1 + B_2, \tag{2}$$

где

$$A_1(x) = (x - \hat{a}_1)' \hat{R}_1^{-1} (x - a_1) \frac{n_1 - m - 1}{n_1 - 1} - \frac{m}{n_1},$$

$$A_2(x) = (x - \hat{a}_2)' \hat{R}_2^{-1} (x - a_2) \frac{n_2 - m - 1}{n_2 - 1} - \frac{m}{n_2},$$

$$B_1(x) = \ln \left\{ \det \hat{R}_1 \cdot \frac{(n_1 - 1)^m}{n_1 (n_1 - 1) \ldots (n_1 - m - 1)} \left(1 - \frac{m}{n_1}\right)^{-1} \right\},$$

$$B_2(x) = \ln \left\{ \det \hat{R}_2 \cdot \frac{(n_2 - 1)^m}{n_2 (n_2 - 1) \ldots (n_2 - m - 1)} \left(1 - \frac{m}{n_2}\right)^{-1} \right\}.$$

Теорема 1. *Пусть* $\lim\limits_{n_i \to \infty} m n_i^{-1} = c_i$, $0 < c_i < \infty$, R_i — *невырожденные матрицы,* x — *нормальный случайный вектор, не зависящий от выборочных значений* $x_i^{(1)}$, $x_i^{(2)}$ *и распределенный по закону* $N(a_1, R_1)$ *либо* $N(a_2, R_2)$. *Тогда при фиксированном векторе* x

$$\operatorname*{plim}_{n_i \to \infty} \{G_F(x) - [(x - a_1)' R_1^{-1} (x - a_1) - (x - a_2)' R_2^{-1} (x - a_2) -$$

$$- \ln \det R_1 + \ln \det R_2]\} = 0.$$

Доказательство. Покажем вначале, что

$$\operatorname*{plim}_{n_1\to\infty} A_1(x) = (x-a_1)' R_1^{-1}(x-a_1). \tag{3}$$

Известно [7], что

$$\hat{a}_1 \approx a_1 + \sqrt{R_1}\nu_1 \frac{1}{\sqrt{n_1}}, \quad \hat{R}_1 \approx \sqrt{R_1}\frac{HH'}{n_1-1}\sqrt{R_1}, \quad \nu_1 \sim N(0, I),$$

$H = (h_{ij})_{i=\overline{1,m},\ j=\overline{1,n_1-1}}$, h_{ij} — независимые $N(0,1)$ респределенные случайные величины. Знак $\approx$ означает совпадение распределений. Легко показать так же, что

$$(x-\hat{a}_1)'\hat{R}_1^{-1}(x-\hat{a}_1) \approx \left(\frac{HH'}{n_1-1}\right)^{-1}_{11}(x-\hat{a}_1)'R_1^{-1}(x-\hat{a}_1),$$

где $(HH')^{-1}_{11}$ — элемент матрицы, обратной к HH'. В общем случае элементы матрицы $(HH')^{-1}$ представимы в виде

$$(HH')^{-1}_{ii} \approx \Big(\sum_{i=m}^{n_1-1}\xi_i^2\Big)^{-1},$$

где ξ_i — независимые $N(0,1)$ распределенные случайные величины. Тогда левая часть выражения (3) может быть преобразована следующим образом:

$$\operatorname*{plim}_{n_i\to\infty}\left\{(x-\hat{a}_1)'\hat{R}_1^{-1}(x-\hat{a}_1)\frac{n_1-m-1}{n_1-1} - \frac{m}{n_1}\right\} =$$

$$= \operatorname*{plim}_{n_i\to\infty}\frac{n_1-1}{\sum\limits_{i=m}^{n_1-1}\xi_i^2}\cdot\frac{n_1-m-1}{n_1-1}\left[(x-a_1)'R_1^{-1}(x-a_1) - \right.$$

$$\left. - 2(x-a_1)'R_1^{-1/2}\nu_1\cdot\frac{1}{\sqrt{n_1}} + \frac{(\nu_1,\nu_1)}{n_1}\right] - \frac{m}{n_1}.$$

Поскольку

$$\operatorname*{plim}_{n_i\to\infty}\Big(\sum_{i=m}^{n_1-1}\xi_i^2\Big)^{-1}(n_1-m-1) = 1,$$

$$\operatorname*{plim}_{n_1\to\infty} 2(x-a_1)'R_1^{-1/2}\nu_1\frac{1}{\sqrt{n_1}} = 0, \quad \operatorname*{plim}_{n_i\to\infty}\frac{(\nu_1,\nu_1)}{n_1} = \frac{m}{c_1},$$

нетрудно видеть, что равенство (3) имеет место.

Аналогично

$$\operatorname*{plim}_{n_2\to\infty}\left\{(x-\hat{a}_2)'\hat{R}_2^{-1}(x-\hat{a}_2)\frac{n_2-m-1}{n_2-1} - \frac{m}{n_2}\right\} = (x-a_2)'R_2^{-1}(x-a_2).$$

Выражения B_1 и B_2 представляют собой G-оценки обобщенных дисперсий $\ln\det R_1$ и $\ln\det R_2$. Их состоятельность доказана в работе [8]. Таким образом, доказана состоятельность оценки $G_F(x)$ в асимптотике Колмогорова, т. е. выражение (2) действительно является G-оценкой квадратичной дискриминантной функции.

Покажем теперь, что оценка $G_F(x)$ является асимптотически нормальной. В работе [8] доказана асимптотическая нормальчость G-оценки обобщенных дисперсий, поэтому в дальнейшем будем рассматривать только первые два слагаемых в выражении (2).

Теорема 2. *Пусть выполняется условие А. Н. Колмогорова*

$$\lim_{n_i\to\infty} mn_i^{-1} = c_i, \quad 0 < c_i < \infty, \quad i = 1, 2,$$

R_1 — *невырожденная матрица и*

$$\operatorname*{plim}_{n_1\to\infty}(a_1-a_2)'R^{-1}(a_1-a_2)\left[\frac{1}{n_1}+\frac{1}{n_2}\right]=0.$$

Тогда

$$\lim_{n_i\to\infty}P\left\{\frac{A_1(x)-(x-a_1)'R_1^{-1}(x-a_1)}{\sqrt{D_m}}\sqrt{n_1-m-1}<z\right\}=$$

$$=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{z}e^{-y^2/2}dy,$$

где $D_m=2\left(\frac{m}{n_1}\right)^2\frac{n_1-m-1}{m}+2\left[(x-a_1)'R_1^{-1}(x-a_1)+\frac{m}{n_1}\right]^2+4(x-$

$-a_1)'R_1^{-1}(x-a_1)\frac{n_1-m-1}{n_1}$.

Доказательство.

$$\{A_1(x)-(x-a_1)'R_1^{-1}(x-a_1)\}\sqrt{n_1-m-1}\approx\left\{\left(\sum_{i=m}^{n_1-1}\frac{\xi_i^2}{n_1-m-1}\right)^{-1}\times\right.$$

$$\times\left[(x-a_1)'R_1^{-1}(x-a_1)+2(x-a_1)'R_1^{-1/2}\nu_1\cdot\frac{1}{\sqrt{n_1}}+\frac{1}{n_1}(\nu_1,\nu_1)\right]-$$

$$\left.-\frac{m}{n_1}-(x-a_1)'R_1^{-1}(x-a_1)\right\}\sqrt{n_1-m-1}.$$

Рассмотрим следующие преобразования:

$$\sum_{i=m}^{n_1-1}\frac{\xi_i^2}{n_1-m-1}=1+\sum_{i=m}^{n_1-1}\frac{\xi_i^2-1}{\sqrt{n_1-m-1}}\cdot\frac{1}{\sqrt{n_1-m-1}},$$

$$\sum_{i=m}^{n_1-1}\frac{\xi_i^2}{n_1-m-1}\approx1+\frac{1}{\sqrt{n_1-m-1}}\cdot\eta_1\sqrt{2},\quad\eta_1\sim N(0,1).$$

Тогда

$$\{A_1(x)-(x-a_1)'R_1^{-1}(x-a_1)\}\sqrt{n_1-m-1}\approx\sqrt{n_1-m-1}\times$$

$$\times\left\{\frac{1}{1+\frac{\eta_1\sqrt{2}}{\sqrt{n_1-m-1}}}\left[(x-a_1)'R_1^{-1}(x-a_1)+\frac{(\nu_1,\nu_1)-m}{m}\frac{m}{n_1}+\right.\right.$$

$$\left.+2(x-a_1)'R_1^{-1/2}\nu_1\frac{1}{\sqrt{n_1}}+\frac{m}{n_1}\right]-\frac{m}{n_1}-(x-a_1)'R_1^{-1}(x-a_1)\Bigg\}=$$

$$=\left\{\left[1+\frac{\frac{\eta_1\sqrt{2}}{\sqrt{n_1-m-1}}}{1+\frac{\eta_1\sqrt{2}}{\sqrt{n_1-m-1}}}\right]\left[(x-a_1)'R_1^{-1}(x-a_1)+\right.\right.$$

$$+2(x-a_1)'R_1^{-1/2}\nu_1\frac{1}{\sqrt{n_1}}+\frac{(\nu_1,\nu_1)-m}{m}\frac{m}{n_1}+\frac{m}{n_1}\Bigg]-\frac{m}{n_1}-(x-$$

$$\left.-a_1)'R_1^{-1}(x-a_1)\right\}\sqrt{n_1-m-1}=\frac{(\nu_1,\nu_1)-m}{\sqrt{m}}\sqrt{\frac{n_2-m-1}{m}}\frac{m}{n_1}+$$

$$+ 2(x-a_1)'R_1^{-1/2}\nu_1 \frac{1}{\sqrt{n_1}} - 1 + \frac{\eta_1\sqrt{2}}{\sqrt{n_1-m-1}}\left[(x-a_1)'R_1^{-1}(x-a_1) + \right.$$

$$\left. + 2(x-a_1)'R_1^{-1/2}\nu_1 \frac{1}{\sqrt{n_1}} + \frac{(\nu_1,\nu_1)-m}{m}\frac{m}{n_1} + \frac{m}{n_1}\right].$$

Очевидно, что

$$\frac{(\nu_1,\nu_1)-m}{\sqrt{m}} = \sum_{i=1}^{m}\frac{(\nu_1^i)^2-1}{\sqrt{m}} \approx \eta_2\sqrt{2}, \tag{4}$$

где ν_1^i — компоненты вектора ν_1, $\eta_2 \sim N(0,1)$; η_1, η_2 — независимые случайные величины,

$$\lim_{n_i\to\infty}\frac{\eta_1\sqrt{2}}{\sqrt{n_1-m-1}} = 0, \tag{5}$$

$$\lim_{n_1\to\infty}\frac{(\nu_1,\nu_1)}{m}\frac{m}{n_1} + \frac{m}{n_1} = \lim_{n_1\to\infty}\frac{(\nu_1,\nu_1)}{m}\frac{m}{n_1} = \frac{m}{n_1}, \tag{6}$$

$$\lim_{n_1\to\infty} 2(x-a_1)'R_1^{-1/2}\nu_1\frac{1}{\sqrt{n_1}} = 0. \tag{7}$$

Используя теперь выражения (4) — (7), получаем

$$\{A_1(x) - (x-a_1)'R_1^{-1}(x-a_1)\}\sqrt{n_1-m-1} \approx$$

$$\approx \eta_2\sqrt{2}\frac{m}{n_1}\sqrt{\frac{n_1-m-1}{m}} - \eta_1\sqrt{2}\left[(x-a_1)'R_1^{-1}(x-a_1) + \frac{m}{n_1}\right] +$$

$$+ 2(x-a_1)'R_1^{-1/2}\nu_1\sqrt{\frac{n_1-m-1}{n_1}}.$$

Легко показать, что если η_i — независимые случайные величины, распределенные по стандартному нормальному закону, а α_i — некоторые константы, то

$$\sum_{i=1}^{n}\alpha_i\eta_i \approx \eta_i\sqrt{\sum_{i=1}^{n}\alpha_i^2}.$$

Тогда

$$\{A_1(x) - (x-a_1)'R_1^{-1}(x-a_1)\}\sqrt{n_1-m-1} \approx$$

$$\approx \eta_2\sqrt{2\frac{m^2}{n_1^2}\frac{n_1-m-1}{m} + 2\left[(x-a_1)'R_1^{-1}(x-a_1) - \frac{m}{n_1}\right]^2 + \ldots \to}$$

$$\leftarrow \ldots \overline{+ 4(x-a_1)'R_1^{-1}(x-a_1)\frac{n_1-m-1}{n_1}},$$

$$\frac{\{A_1(x) - (x-a_1)'R_1^{-1}(x-a_1)\}}{\sqrt{D_m}}\sqrt{n_1-m-1} \approx \eta_2,$$

что и доказывает утверждение теоремы 2.

Аналогично георему 2 можно показать, что

$$\lim_{n_2\to\infty} P\left\{\frac{A_2(x) - (x-a_2)'R_2^{-1}(x-a_2)}{\sqrt{k_m}}\sqrt{n_2-m-1} < z\right\} =$$

$$= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-y^2/2}dy,$$

где

$$k_m = 2\left(\frac{m}{n_2}\right)^2 \frac{n_2 - m - 1}{m} + 2\left[(x - a_2)' R_2^{-1} (x - a_2) + \frac{m}{n_2}\right]^2 +$$

$$+ 4 (x - a_2)' R_2^{-1} (x - a_2) \frac{n_2 - m - 1}{n_2}.$$

Таким образом, построенная оценка $G_F(x)$ является асимптотически нормальной.

1. *Гирко В. Л.* «Борьба с размерностью» в многомерном статистическом анализе // Тез. третьей Всесоюз. науч-техн. конф. «Применение многомерного статистического анализа в экономике и оценки качества продукции».— Тарту, 1985.— С. 43—52.
2. *Деев А. Д.* Представление статистик дискриминантного анализа и асимптотические разложения при размерности пространства, сравнимой с объемом выборок // Докл. АН СССР.— 1970.— **195**, № 4.— С. 759—762.
3. *Мешалкин Л. Д., Сердобольский В. И.* Ошибки при классификации многомерных наблюдений // Теория вероятностей и ее применения.— 1978.— 23, вып. 4.— С. 772—781.
4. *Троицкий Е. В.* Исследование квадратичного дискриминатора при большом числе признаков.— М., 1979.— 24 с.— Деп. в ВИНИТИ, № 4384-79 Деп.
5. *Фукунага К.* Введение в статистическую теорию распознавания образов.— М. : Наука, 1979.— 368 с.
6. *Гирко В. Л.* G-анализ наблюдений большой размерности // Вычислит. и прикл. математика.— 1986.— Вып. 60.— С. 115—121.
7. *Андерсон Т.* Введение в многомерный статистический анализ.— М. : Физматгиз, 1963.— 500 с.
8. *Гирко В. Л.* Введение в общий статистический анализ // Теория вероятностей и ее применения.— 1987.— **32**, вып. 2.— С. 252—265.

Киев. ун-т

Получено 08.01.88