Записки научных
семинаров ПОМИ
Том 466, 2017 г.

И. А. Суслина, О. В. Соколов

СВЯЗЬ ЗАДАЧИ СЕЛЕКЦИИ РАЗРЕЖЕННОЙ ПОДМАТРИЦЫ МАТРИЦЫ БОЛЬШОГО РАЗМЕРА И БАЙЕСОВСКОЙ ЗАДАЧИ ПРОВЕРКИ ГИПОТЕЗ

§1. Введение

Пусть наблюдения представлены в виде матрицы $\mathbf{Y} = \{Y_{ij}\}$ размера $N \times M$

$$Y_{ij} = s_{ij} + \xi_{ij}, \quad i = 1, \dots, N, \quad j = 1, \dots, M, \tag{1.1}$$

где $\{\xi_{ij}\}$ – независимые одинаково распределенные случайные величины и $s_{ij} \geqslant 0$, для всех $i \in \{1, \dots, N\}$, $j \in \{1, \dots, M\}$. Ошибки наблюдения ξ_{ij} распределены по стандартному гауссовскому закону.

Матрица размера $N \times M$ содержит $L = C_N^n \cdot C_M^m$ различных подматриц $\mathcal{C} = \mathcal{C}_k$, $k = 1, \dots, L$ размера $n \times m$. Пусть

$$C_k = \big\{C = A \times B \subset \{1, \dots, N\} \times \{1, \dots, M\}, \\ \operatorname{Card}(A) = n, \ \operatorname{Card}(B) = m\big\}$$

представляет из себя набор индексов (i, j) элементов подматрицы $\mathcal{C}_k$, $k = 1, \dots, L$.

Рассмотрим $(L + 1)$ простую гипотезу:

$$\begin{aligned} H_0 :& \quad Y_{ij} = \xi_{ij} \quad \text{при любых } (i, j), \quad i \in \{1, \dots, N\}, \quad j \in \{1, \dots, M\}, \\ H_k :& \quad Y_{ij} = \begin{cases} \xi_{ij} & \text{при } (i, j) \notin C_k, \\ \xi_{ij} + s_{ij}, \ s_{ij} \geqslant a > 0 & \text{при } (i, j) \in C_k, \end{cases} \\ & \quad k = 1, \dots, L. \end{aligned} \tag{1.2}$$

То есть, гипотеза H_0 состоит в том, что все элементы матрицы $\mathbf{Y}$ центрированы, а гипотеза H_k – в том, что центрированы все элементы, кроме элементов подматрицы $\mathcal{C}_k$ размера $n \times m$, а именно: $s_{ij} \geqslant a > 0$ для пар (i, j) из множества индексов C_k и $s_{ij} = 0$ для пар $(i, j) \notin C_k$.

Таким образом, перед нами стоит задача проверки гипотез, в частности, нахождения такого теста, с помощью которого можно будет по

Ключевые слова: байесовская задача проверки гипотез, байесовский риск, минимаксный риск, селекция, оптимальный байесовский тест, состоятельный тест.

исходным наблюдениям, представленным в виде матрицы $\mathbf{Y}$, определить, какая гипотеза будет принята. Качество теста характеризуется байесовским и минимаксным рисками.

Измеримые функции, определенные на множестве наблюдений $\{\mathbf{Y}\}$ со значениями в множестве $\{0, 1, \ldots, L\}$, назовем тестами:

$$\psi = \psi(\mathbf{Y}) : \{\mathbf{Y}\} \to \{0, \ldots, L\}.$$

Пусть $\mathbf{P}_{H_k}$ – мера, соответствующая гипотезе H_k. Для каждого теста ψ, вводя обозначения $R_k(\psi) = \mathbf{P}_{H_k}(\psi \neq k)$, определим его риск. Положим

$$R(\psi) = \max_{k=0,\ldots,L} \mathbf{P}_{H_k}\big(\psi(\mathbf{Y}) \neq k\big) = \max_{k=0,\ldots,L} \big\{R_k(\psi)\big\}.$$

Минимаксный риск определяется соотношением

$$R_M = \inf_{\psi} R(\psi),$$

где инфимум берется по всем измеримым отображениям $\{\mathbf{Y}\}$ в $\{0, 1, \ldots, L\}$.

Нас будут интересовать асимптотические свойства минимаксного риска в случае разреженных подматриц $\mathcal{C}_k = \mathcal{C}_k(N, M, n, m, a)$. Предположим, что

$$\begin{aligned} &N \to \infty, \quad M \to \infty, \quad n \to \infty, \quad m \to \infty, \\ &p = n/N \to 0, \quad q = m/M \to 0. \end{aligned} \tag{1.3}$$

Назовем тест состоятельным в минимаксном смысле, если $R(\psi) \to 0$ при выполнении условий (1.3).

Будем искать граничные значения для a: нижние границы – условия, при которых минимаксный риск стремится к единице, а также верхние границы – условия, при которых минимаксный риск стремится к нулю. При изучении верхних границ будет построен состоятельный в минимаксном смысле тест.

Аналогичные задачи в несколько иной постановке изучались, в частности в [1–5]. Мы существенно используем результаты работы [1]. В ней достаточно подробно представлена история изучения подобных задач. Результаты, полученные для гауссовского случая в работе [2], верны только по порядку. Поэтому результаты, приведенные в ней для случая, когда ошибки наблюдения ξ_{ij} распределены не по стандартному гауссовскому закону, не представляется возможным использовать

в наших рассмотрениях. В работе [3] показано, что проверка гипотезы

$$H_0: \quad Y_{ij} = \xi_{ij} \quad \text{при любых} \ \ (i,j), \ \ i \in \{1,\dots,N\}, \quad j \in \{1,\dots,M\},$$

против альтернативы: существует такая подматрица $\mathcal{C} = \mathcal{C}_k$, $k = 1,\dots,L$, размера $n \times m$ исходной матрицы $\mathbf{Y} = \{Y_{ij}\}$ размера $N \times M$, что

$$Y_{ij} = \begin{cases} \xi_{ij} & \text{при } (i,j) \notin C_k, \\ \xi_{ij} + s_{ij}, \ s_{ij} \geqslant a > 0 & \text{при } (i,j) \in C_k, \end{cases}$$

возможна в более широких условиях, чем изучаемая нами задача (см. [1], глава 2.2).

§2. Оптимальный байесовский тест

Рассмотрим байесовскую задачу проверки k простых гипотез. Пусть $\{\mathcal{X}, \mathcal{A}, \mathcal{P}\}$ – статистический эксперимент и пусть $\mathbf{P}_1,\dots,\mathbf{P}_k$ – вероятностные меры, $\mathbf{P}_l \in \mathcal{P}$ соответствует l-ой гипотезе о наблюдении $x \in \mathcal{X}$. Пусть $\pi = \{\pi_1,\dots,\pi_k\}$ – априорные вероятности соответствующих гипотез, где $\pi_l \geqslant 0, \ l = 1,\dots,k, \ \sum\limits_{l=1}^{k} \pi_l = 1$. Для теста $\psi = \psi(x)$ – измеримого отображения $\mathcal{X} \to \{1,\dots,k\}$ рассмотрим вероятности ошибок $R_l(\psi) = \mathbf{P}_l(\psi \neq l) = 1 - \mathbf{P}_l(\psi = l)$. Их среднее значение (смесь по мере π) назовем байесовским риском теста:

$$R_B(\psi) = \sum_{l=1}^{k} \pi_l \, R_l(\psi) = 1 - \sum_{l=1}^{k} \pi_l \, \mathbf{P}_l(\psi = l).$$

Будем называть тест $\psi^*(x)$ *оптимальным*, если $R_B(\psi^*) = \inf_\psi R_B(\psi) = R_B$, где инфимум берется по всем измеримым отображениям $\mathcal{X}$ в $\{1,\dots,k\}$. Хорошо известно, что для любой априорной меры $\pi = \{\pi_0,\dots,\pi_L\}$, $\sum\limits_{k=0}^{L} \pi_k = 1$, байесовский риск не превосходит минимаксного риска. То есть,

$$R_B = \inf_\psi \sum_{k=0}^{L} \pi_k \, R_k(\psi) \leqslant R_M = \inf_\psi R(\psi) = \inf_\psi \max_{k=0,\dots,L} R_k(\psi). \qquad (2.1)$$

Будем предполагать, что при всех $l = 1,\dots,k$ существуют $f_l(x)$ – плотности мер $\mathbf{P}_l$ по отношению к общей мере $\mathbf{P}$.

Предложение 2.1. *Оптимальный тест в байесовской проверке k простых гипотез имеет вид:*

$$\psi^*(x) = \operatorname*{argmax}_{1 \leqslant l \leqslant k} \pi_l \, f_l(x).$$

Таким образом, тест $\psi^*(x) = l_0$, если $\pi_{l_0} \, f_{l_0}(x) \geqslant \pi_l \, f_l(x)$ при $l \in \{1, \ldots, L\}$, $l \neq l_0$. В случае равенства $\pi_{l_0} \, f_{l_0}(x) = \pi_l \, f_l(x)$ при $l \in \{1, \ldots, L\}$, $l \neq l_0$, применяется дополнительная процедура розыгрыша.

Доказательство. Нужно найти минимум следующего выражения:

$$R_B(\psi) = \sum_{i=1}^{i=k} \mathbf{P}_i(\psi \neq i) \, \pi_i,$$

где риск минимизируется по множеству всех тестов $\psi : \mathcal{X} \to \{1, \ldots, k\}$. Имеем

$$R_B(\psi) = \sum_{i=1}^{i=k} \mathbf{P}_i(\psi \neq i) \, \pi_i = 1 - \sum_{i=1}^{i=k} \mathbf{P}_i(\psi = i) \, \pi_i.$$

Таким образом, задача сводится к нахождению максимума выражения:

$$W = \sum_{i=1}^{i=k} \mathbf{P}_i(\psi = i) \, \pi_i. \tag{2.2}$$

В силу того, что $f_i(x)$ является плотностью вероятностной меры $\mathbf{P}_i$, выражение (2.2) примет вид:

$$W = \sum_{i=1}^{i=k} \int\limits_{\mathcal{X}_i} \pi_i \, f_i(x) \, dx, \tag{2.3}$$

где $\mathcal{X}_i = \{x \in \mathcal{X} : \ \psi(x) = i\}$, $i = 1, \ldots, k$. Легко видеть, что сумма (2.3) достигает максимума, если

$$\mathcal{X}_i = \big\{ \, x \in \mathcal{X} : \quad \pi_i \, f_i(x) \geqslant \pi_j \, f_j(x), \ \ \forall j = 1, \ldots, k \, \big\}. \qquad \square$$

Замечание 2.1. Если мы хотим связать байесовскую задачу проверки гипотез (1.2) с селекцией, то естественно предполагать, что априорные вероятности π_k, $k = 1, \ldots, L$, совпадают.

Предложение 2.2. *Обозначим $M = \dfrac{mna}{2} + \dfrac{\log L}{a} + \dfrac{\log \frac{p}{1-p}}{a}$. Рассмотрим априорную меру $\pi = (\pi_0, \ldots, \pi_L)$, $\pi_0 = p$, $\pi_k = \frac{1-p}{L}$, $k = 1, \ldots, L$,*

где $0 < p < 1$ *– вероятность того, что* $s_{ij} = 0$ *для любых* i, j*. Тогда тест*

$$\tilde{\psi}^{\mathrm{op}}(\mathbf{Y}) = \begin{cases} 0, & \max\limits_k \Big\{ \sum\limits_{C_k} Y_{ij} \Big\} < M, \\ l > 0, & \max\limits_k \Big\{ \sum\limits_{C_k} Y_{ij} \Big\} \geqslant M, \ \operatorname{argmax}\limits_k \Big\{ \sum\limits_{C_k} Y_{ij} \Big\} = C_l, \end{cases} \tag{2.4}$$

является оптимальным байесовским тестом для проверки гипотез (1.2)*, если* $s_{ij} = 0$ *или* $s_{ij} = a$*.*

Доказательство. Из предложения 2.1 следует, что оптимальный байесовский тест в данной задаче должен иметь вид:

$$\psi(\mathbf{Y}) = \operatorname*{argmax}_{0 \leqslant k \leqslant L} \pi_k \, f_k(Y),$$

где $f_0(Y)$ соответствует матрице $\mathbf{Y}$ с центрированными элементами $Y_{ij} \sim N(0,1)$, а каждая функция $f_k(Y)$, $k = 1, \ldots, L$, соответствует матрице $\mathbf{Y}$, в которой $s_{ij} = 0$ при $(i,j) \notin C_k$ и $s_{ij} = a > 0$ при $(i,j) \in C_k$. Аргумент максимума не изменится, если разделить все выражения на $f_0(Y)$ – положительную константу. Следовательно, оптимальный тест можно искать в виде:

$$\psi(\mathbf{Y}) = \operatorname*{argmax}_{0 \leqslant k \leqslant L} \pi_k \, \frac{f_k(Y)}{f_0(Y)},$$

где $\dfrac{f_k(Y)}{f_0(Y)} = \dfrac{d\mathbf{P}_k}{d\mathbf{P}_0}$ – отношение правдоподобия для гипотезы H_k. Здесь $f_0(Y)$ соответствует гипотезе $H_0 : s_{ij} = 0$ при всех допустимых (i,j), а $f_k(Y)$ соответствует гипотезе $H_k : s_{ij} = a > 0$, $(i,j) \in C_k$, $s_{ij} = 0$, $(i,j) \notin C_k$.

Для H_0 отношение правдоподобия равно 1, найдем $\dfrac{f_k(Y)}{f_0(Y)}$ при $k = 1, \ldots, L$. Все $\{\xi_{ij}\}$ независимы и распределены по стандартному нормальному закону. Следовательно,

$$\frac{f_k(Y)}{f_0(Y)} = \frac{\prod\limits_{(i,j) \notin C_k} \exp\big(-Y_{ij}^2/2\big) \times \prod\limits_{(i,j) \in C_k} \exp\big(-(Y_{ij} - a)^2/2\big)}{\prod\limits_{\substack{1 \leqslant i \leqslant N, \\ 1 \leqslant j \leqslant M}} \exp\big(-Y_{ij}^2/2\big)}.$$

Приходим к выражению:

$$\frac{f_k(Y)}{f_0(Y)} = \exp\Big(-\frac{a^2nm}{2} + a\sum_{(i,j)\in C_k} Y_{ij}\Big).$$

Откуда

$$\pi_0\,\frac{f_0(Y)}{f_0(Y)} = p,$$
$$\pi_k\,\frac{f_k(Y)}{f_0(Y)} = \frac{1-p}{L}\,\exp\Big(-\frac{a^2nm}{2} + a\sum_{(i,j)\in C_k} Y_{ij}\Big), \quad k = 1,\dots,L.$$

В результате использования оптимального байесовского теста гипотеза H_0 принимается при условии:

$$\max_k \sum_{C_k} Y_{ij} < \frac{mna}{2} + \frac{\log(L)}{a} + \frac{\log\dfrac{p}{1-p}}{a}.$$

Если это условие не выполнено, в выражении $\pi_k\,\dfrac{f_k(Y)}{f_0(Y)}$ нужно максимизировать сумму $\sum\limits_{(i,j)\in C_k} Y_{ij}$. Следовательно, тест должен принимать гипотезу H_l, если $\operatorname{argmax}\limits_k \sum\limits_{C_k} Y_{ij} = C_l$. Таким образом, тест (2.4) является оптимальным байесовским тестом. □

Заметим, что при выполнении условий (1.3)

$$\frac{mna}{2} + \frac{\log(L)}{a} + \frac{\log\dfrac{p}{1-p}}{a} \sim \frac{mna}{2} + \frac{\log(L)}{a} = M_1.$$

Далее мы будем рассматривать и анализировать тест

$$\psi^{\mathrm{op}}(\mathbf{Y}) = \begin{cases} 0, & \max\limits_k\Big\{\sum\limits_{C_k} Y_{ij}\Big\} < M_1, \\ l > 0, & \max\limits_k\Big\{\sum\limits_{C_k} Y_{ij}\Big\} \geqslant M_1,\ \operatorname{argmax}\limits_k\Big\{\sum\limits_{C_k} Y_{ij}\Big\} = C_l. \end{cases} \tag{2.5}$$

Замечание 2.2. Заметим, что из соображений симметрии $R_i(\psi^{\mathrm{op}}) = R_j(\psi^{\mathrm{op}})$ при $1 \leqslant i, j \leqslant L$.

§3. Верхние границы

Пусть выполнено (1.3). Укажем условия, при которых минимаксный риск $R_M \to 0$, а также покажем, что при этих условиях тест (2.5) является состоятельным, то есть, его минимаксный риск стремится к нулю. Заметим, что, если $R_M \to 0$, то байесовский риск $R_B = \inf_\psi R_B(\psi) \to 0$ для любого набора априорных вероятностей $\pi = \{\pi_1, \dots, \pi_k\}$, $p_i \geqslant 0$, $\sum_{i=1}^k p_i = 1$.

Определим основные величины, необходимые в дальнейших рассмотрениях:

$$\begin{gathered} B = B_{n,\,m,\,N,\,M} = \min\{A_1,\, A_2,\, A\} \\ \text{где} \quad A = \frac{a\sqrt{nm}}{\sqrt{2(n\log(p^{-1}) + m\log(q^{-1}))}}, \\ A_1 = \frac{a\sqrt{m}}{\sqrt{2}(\sqrt{\log(n)} + \sqrt{\log(N-n)})}, \\ A_2 = \frac{a\sqrt{n}}{\sqrt{2}(\sqrt{\log(m)} + \sqrt{\log(M-m)})}. \end{gathered} \tag{3.1}$$

Нам потребуется следующее замечание.

Замечание 3.1. Функция распределения суммы $Y = \sum_{i=1}^n X_i$ – независимых случайных величин, распределенных по стандартному нормальному закону, имеет вид $F_Y(t) = \mathbf{P}(Y < t) = \Phi(t/\sqrt{n})$, где $\Phi(\cdot)$ – функция распределения стандартного нормального закона. Кроме того, при выполнении условий (1.3) справедливо соотношение $\log(C_N^n) \sim n\log(N/n)$.

Предложение 3.1. *Пусть выполнено* (1.3) *и* $B = B_{n,\,m,\,N,\,M}$, *определенное* (3.1), *удовлетворяет условию*

$$\liminf B_{n,\,m,\,N,\,M} > 1. \tag{3.2}$$

Тогда тест, определенный (2.5), *является состоятельным, то есть* $R(\psi^{op}) = \max\limits_{k=0,\dots,L} \mathbf{P}_{H_k}(\psi(\mathbf{Y}) \neq k) \to 0$.

Доказательство. Так как

$$R(\psi^{\mathrm{op}}) = \max_{k=0,\dots,L} R_k(\psi^{\mathrm{op}}) = \max(R_0(\psi^{\mathrm{op}}), R_1(\psi^{\mathrm{op}})),$$

нужно доказать, что $R_0(\psi^{\mathrm{op}})$ и $R_1(\psi^{\mathrm{op}})$ одновременно стремятся к нулю.

I. **Оценка риска** $R_1(\psi^{\mathrm{op}})$**.** Покажем, что при выполнении условий предложения $R_1(\psi^{\mathrm{op}}) \to 0$. Очевидно, что

$$\begin{aligned} R_1(\psi^{\mathrm{op}}) &= \mathbf{P}_{H_1}(\psi^{\mathrm{op}} \neq 1) = \mathbf{P}_{H_1}\Big(\max_k \sum_{C_k} Y_{ij} > \sum_{C_1} Y_{ij}\Big) \\ &\quad + \mathbf{P}_{H_1}\Big(\max_k \sum_{C_k} Y_{ij} = \sum_{C_1} Y_{ij},\ \sum_{C_1} Y_{ij} < \frac{mna}{2} + \frac{\log L}{a}\Big) \qquad (3.3) \\ &\leqslant \mathbf{P}_1 + \mathbf{P}_2, \end{aligned}$$

где $\mathbf{P}_1 = \mathbf{P}_{H_1}\Big(\max\limits_k \sum\limits_{C_k} Y_{ij} > \sum\limits_{C_1} Y_{ij}\Big)$, $\mathbf{P}_2 = \mathbf{P}_{H_1}\Big(\sum\limits_{C_1} Y_{ij} < \dfrac{mna}{2} + \dfrac{\log L}{a}\Big)$.

В доказательстве теоремы 2.1 в [1] показывается, что $\mathbf{P}_1 \to 0$ при выполнении условий (1.3) и (3.2). Значит это верно и в условиях предложения.

Рассмотрим теперь $\mathbf{P}_2$ – второе слагаемое в последнем выражении (3.3). Для него справедливо равенство

$$\begin{aligned} \mathbf{P}_2 &= \mathbf{P}_{H_1}\Big(\sum_{C_1} Y_{ij} < \frac{mna}{2} + \frac{\log L}{a}\Big) = \mathbf{P}_{H_1}\Big(\sum_{C_1} (\xi_{ij} + a) < \frac{mna}{2} + \frac{\log L}{a}\Big) \\ &= \mathbf{P}_{H_1}\Big(\sum_{C_1} \xi_{ij} + mna < \frac{mna}{2} + \frac{\log L}{a}\Big) \\ &= \mathbf{P}_{H_1}\Big(\sum_{C_1} \xi_{ij} < -\frac{mna}{2} + \frac{\log L}{a}\Big). \end{aligned}$$

Используя замечание 3.1, имеем

$$\begin{aligned} \mathbf{P}_{H_1}\Big(\sum_{C_1} \xi_{ij} < -\frac{mna}{2} + \frac{\log L}{a}\Big) &= \Phi\Big(-\frac{a\sqrt{mn}}{2} + \frac{\log L}{a\sqrt{mn}}\Big) \\ &= \Phi\Big(-\frac{a\sqrt{mn}}{2}\Big(1 - \frac{2\log L}{a^2 mn}\Big)\Big). \end{aligned}$$

Кроме того, снова используя замечание 3.1, условие (3.2), а также равенство $C_N^n \cdot C_M^m = L$, получаем

$$\frac{a^2 mn}{2\log L} \sim 1 + \delta, \quad \delta > 0, \qquad (3.4)$$

где величина $\delta = \delta(N, M, n, m, a)$ равномерно отделена от нуля при выполнении (1.3). Следовательно,

$$\Phi\Big(-\frac{a\sqrt{mn}}{2}\Big(1-\frac{2\log L}{a^2mn}\Big)\Big) \sim \Phi\Big(-\frac{a\sqrt{mn}}{2}\frac{\delta}{(1+\delta)}\Big).$$

Из условия (3.2) следует, что $a\sqrt{mn} \to \infty$. Отсюда получаем, что $\mathbf{P}_2 \to 0$, условия (1.3), (3.2) влекут $R_1(\psi^{\mathrm{op}}) \to 0$.

II. **Оценка риска R_0.** Используя хорошо известную асимптотику

$$\Phi(x) \sim \frac{1}{\sqrt{2\pi}}\frac{e^{-x^2/2}}{(-x)} \quad \text{при } x \to -\infty, \tag{3.5}$$

покажем, что $R_0(\psi^{\mathrm{op}}) \to 0$. Имеем

$$\begin{aligned} R_0(\psi^{\mathrm{op}}) =& \mathbf{P}_{H_0}\big(\psi^{\mathrm{op}} \neq 0\big) = \mathbf{P}_{H_0}\Big(\max_{k=1,\dots,L}\sum_{(i,j)\in C_k} Y_{ij} \geqslant \frac{mna}{2} + \frac{\log L}{a}\Big) \\ &\leqslant \sum_{k=1}^{L} \mathbf{P}_{H_0}\Big(\sum_{(i,j)\in C_k} Y_{ij} \geqslant \frac{mna}{2} + \frac{\log L}{a}\Big) \\ &= L\cdot\Phi\Big(-\Big(\frac{a\sqrt{mn}}{2} + \frac{\log L}{a\sqrt{mn}}\Big)\Big). \end{aligned}$$

Из (3.5) следует, что при выполнении (1.3) для достаточно больших n, m справедливо неравенство:

$$L\cdot\Phi\Big(-\Big(\frac{a\sqrt{mn}}{2} + \frac{\log L}{a\sqrt{mn}}\Big)\Big) \leqslant \exp\Big(-\Big(\frac{a\sqrt{mn}}{2} + \frac{\log L}{a\sqrt{mn}}\Big)^2\Big/2 + \log L\Big). \tag{3.6}$$

Рассмотрим правую часть неравенства (3.6). Из соотношения (3.4) получаем

$$\begin{aligned} \frac{mna^2}{8} - \frac{\log L}{2} + \frac{\log^2 L}{2mna^2} &\sim \frac{1+\delta}{4}\log L - \frac{\log L}{2} + \frac{\log L}{4(1+\delta)} \\ &= \frac{\log L}{4}\frac{\delta^2}{1+\delta} \to \infty. \end{aligned}$$

Следовательно, в условиях предложения 3.1, имеем $R_0(\psi^{\mathrm{op}}) \to 0$. $\square$

Следствие 3.1. *При выполнении условий* (1.3), (3.2) *минимаксный риск* $R_M = \inf_\psi R(\psi) \leqslant R(\psi^{\mathrm{op}}) \to 0$. *Обозначим через* $R_B(\pi)$ *байесовский риск, возникающий при априорном распределении* π. *Имеем*

$R_B(\pi) \leqslant R_M$, следовательно, $R_B(\pi) \to 0$ при любом априорном распределении π.

Замечание 3.2. В параграфе 3 все рассмотрения и доказательства проводились в предположении, что $s_{ij} = 0$ или $s_{ij} = a$. Легко видеть (сравни с [1], замечание 2.1), что все рассмотрения и доказательства переносятся на случай $s_{ij} = 0$ или $s_{ij} \geqslant a$.

§4. Нижние границы

Укажем условия, при которых минимаксный риск R_M стремится к 1.

Предложение 4.1. *Пусть выполнено* (1.3) *и* $B = B_{n,m,N,M}$, *определенное* (3.1), *удовлетворяет условию*

$$\limsup B_{n,m,N,M} < 1. \tag{4.1}$$

Тогда

$$\mathbf{P}_{H_1}\big(\max_k \sum_{C_k} Y_{ij} > \sum_{C_1} Y_{ij}\big) \to 1.$$

Доказательство. Рассмотрим задачу чистой селекции, которая аналогична данной, но исключает возможность реализации гипотезы H_0. То есть, в матрице $\mathbf{Y}$ обязательно есть подматрица C_k, такая что $s_{ij} = a > 0$ для $(i,j) \in C_k$ и $s_{ij} = 0$ для $(i,j) \notin C_k$. Теорема 2.2 в [1] утверждает, что при выполнении условий (1.3), (4.1)

$$\inf_\psi \big\{ \max_{k=1,\dots,L} \mathbf{P}_{H_k}\big(\psi(\mathbf{Y}) \neq k\big)\big\} \to 1,$$

где инфимум берется по всем измеримым отображениям $\{\mathbf{Y}\}$ в множество $\{1,\dots,L\}$.

Рассмотрим тест

$$\psi^+(\mathbf{Y}) = \operatorname*{argmax}_{k=1,\dots,L} \Big\{ \sum_{(i,j)\in C_k} Y_{ij} \Big\}.$$

В силу симметрии $\mathbf{P}_{H_i}\big(\psi^+(\mathbf{Y}) \neq i\big) = \mathbf{P}_{H_j}\big(\psi^+(\mathbf{Y}) \neq j\big)$, $1 \leqslant i,j \leqslant L$, следовательно,

$$\begin{aligned}\inf_\psi \big\{ \max_{k=1,\dots,L} \mathbf{P}_{H_k}\big(\psi(\mathbf{Y}) \neq k\big)\big\} &\leqslant \max_{k=1,\dots,L} \mathbf{P}_{H_k}\big(\psi^+(\mathbf{Y}) \neq k\big) \\ &= \mathbf{P}_{H_1}\big(\psi^+(\mathbf{Y}) \neq 1\big) = \mathbf{P}_{H_1}\big(\max_k \sum_{C_k} Y_{ij} > \sum_{C_1} Y_{ij}\big). \quad \square\end{aligned}$$

Предложение 4.2. *Обозначим* $R = mna/2$. *Тест*

$$\psi^u(\mathbf{Y}) = \begin{cases} 0, & \max\limits_{k\in\{1,\dots,L\}}\Big\{\sum\limits_{C_k} Y_{ij}\Big\} \leqslant R, \\ l, & \operatorname{argmax}\limits_{k}\Big\{\sum\limits_{C_k} Y_{ij}\Big\} = C_l, \ \max\limits_{k\in\{1,\dots,L\}}\Big\{\sum\limits_{C_k} Y_{ij}\Big\} > R, \end{cases}$$

является оптимальным байесовским тестом при априорной мере $\pi = (\pi_0, \dots, \pi_L)$, $\pi_k = 1/(L+1)$, $k = 0, \dots, L$.

Доказательство. Из предложения 2.1 следует, что оптимальный байесовский тест в данной задаче должен иметь вид:

$$\psi(\mathbf{Y}) = \operatorname*{argmax}_{0\leqslant k\leqslant L} \pi_k f_k(\mathbf{Y}),$$

где $f_0(Y)$ соответствует матрице $\mathbf{Y}$ с центрированными элементами $Y_{ij} \sim N(0,1)$, а $f_k(Y)$, $k = 1, \dots, L$, соответствует матрице $\mathbf{Y}$, в которой $\mathbf{E}(Y_{ij}) = 0$ при $(i,j) \notin C_k$ и $\mathbf{E}(Y_{ij}) = a > 0$ при $(i,j) \in C_k$.

Следовательно, оптимальный тест можно искать в виде:

$$\psi(\mathbf{Y}) = \operatorname*{argmax}_{0\leqslant k\leqslant L} \frac{f_k(Y)}{f_0(Y)},$$

где $\dfrac{f_k(Y)}{f_0(Y)} = \dfrac{d\mathbf{P}_k}{d\mathbf{P}_0}$ – отношение правдоподобия для гипотезы H_k. Для H_0 отношение правдоподобия равно единице, отношения $\dfrac{f_k(Y)}{f_0(Y)}$ для $k = 1, \dots, L$ были найдены при доказательстве предложения 2.2.

Имеем

$$\frac{f_0(Y)}{f_0(Y)} = 1, \tag{4.2}$$

$$\frac{f_k(Y)}{f_0(Y)} = \exp\Big(-\frac{a^2nm}{2} + a \sum_{(i,j)\in C_k} Y_{ij}\Big), \quad k = 1, \dots, L. \tag{4.3}$$

Легко видеть, что $\dfrac{f_0(Y)}{f_0(Y)} = 1$ является максимумом при условии

$$\sum_{C_k} Y_{ij} < R, \quad k \in \{1, \dots, L\},$$

то есть, в этом случае оптимальный тест должен принимать значение 0. При

$$\max_{k\in\{1,\dots,L\}}\Big\{\sum_{(i,j)\in C_k} Y_{ij}\Big\} > R$$

максимальное отношение правдоподобия соответствует максимальной сумме $\sum\limits_{C_k} Y_{ij}$. При $\max\limits_{k\in\{1,\dots,L\}}\Big\{\sum\limits_{C_k} Y_{ij}\Big\} = R$ максимальное отношение правдоподобия достигается по крайней мере при двух гипотезах. Будем считать, что в этом случае принимается H_0. □

Предложение 4.3. *Пусть выполнены условия* (1.3) *и* (4.1). *Тогда минимаксный риск* $R_M \to 1$.

Доказательство. Рассмотрим априорную меру $\pi = (\pi_0, \dots, \pi_L)$, $\pi_k = 1/(L+1), k = 0, \dots, L$. Из предложения 4.2 следует, что оптимальным байесовским тестом при данной априорной мере является тест ψ^u и $R_B(\psi^u) = R_B$. Согласно (2.1) $R_B(\psi^u) = R_B \leqslant R_M$. Следовательно, достаточно доказать, что в условиях предложения $R_B(\psi^u) \to 1$. Оценим байесовский риск теста ψ^u:

$$R_B(\psi^u) = \sum_{k=0}^{L} \frac{1}{L+1} R_k(\psi^u) \geqslant \frac{L}{L+1} R_1(\psi^u). \tag{4.4}$$

Из предложения 4.1 следует

$$R_1(\psi^u) = \mathbf{P}_{H_1}(\psi^u \neq 1) \geqslant \mathbf{P}_{H_1}\Big(\max_k \sum_{C_k} Y_{ij} > \sum_{C_1} Y_{ij}\Big) \to 1. \tag{4.5}$$

□

Литература

1. C. Butucea, Y. I. Ingster, I. A. Suslina, *Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix.* — ESAIM: Probab. Statist. **19** (2015), 115–134.
2. M. Kolar, S. Balakrishnan, A. Rinaldo, A. Singh, *Minimax localization of structural information in large noisy matrices.* — In: Advances in Neural Information Processing Systems 24, J. Shawe-Taylor et al. eds., pp. 909–917, NIPS, 2011.
3. C. Butucea, Y. I. Ingster, *Detection of a sparse submatrix of a high-dimensional noisy matrix.* — Bernoulli **19**, No. 5B (2013), 2652–2688.
4. C. Butucea, G. Gayraud, *Sharp detection of smooth signals in a high-dimensional sparse matrix with indirect observations.* — Ann. Inst. H. Poincaré: Probab. Statist. **52**, No. 4 (2016), 1564–1591.

5. X. Sun, A. B. Nobel, *On the maximal size of large-average and ANOVA-fit submatrices in a Gaussian random matrix.* — Bernoulli **19**, No. 1 (2013), 275–294.

Suslina I. A., Sokolov O. V. The connection between the selection problem for a sparse submatrix of a large-size matrix and the Bayes problem of hypothesis testing.

We associate the selection problem for a sparse submatrix of a matrix of large dimension and the problem of testing the hypothesis of the existence of a sparse submatrix possessing the required properties with the Bayesian hypothesis testing problem.

Поступило 8 ноября 2017 г.

С.-Петербургский Национальный
исследовательский Университет
информационных технологий,
механики и оптики,
Кронверкский проспект, дом 49,
197101, Санкт-Петербург, Россия
E-mail: isuslina@mall.ru

МакКинзи и Компания
улица Лесная, дом 5, строение "С",
125047, Москва, Россия
E-mail: olesokolov@gmail.com