Записки научных
семинаров ПОМИ
Том 486, 2019 г.

А. А. Липатьев, В. В. Ульянов

НЕАСИМПТОТИЧЕСКИЙ АНАЛИЗ СТАТИСТИКИ ЛОУЛИ–ХОТЕЛЛИНГА ДЛЯ ДАННЫХ БОЛЬШОЙ РАЗМЕРНОСТИ

§1. Введение

В большом числе прикладных задач исследователи анализируют многомерные данные, в которых количество p признаков сравнимо с числом n наблюдений. Данные такого рода встречаются в медицине и биологии (к примеру, данные ДНК-микрочипов могут содержат большое количество признаков). Также подобная структура данных встречается при анализе информации из социальных сетей, в области финансов и т.д.

Для анализа данных фиксированной размерности существует множество статистических процедур, уже ставших классическими. Однако часто нет возможности использовать традиционную статистическую процедуру, лишь устремив в ней количество признаков к бесконечности, так как при этом изменяется предельное распределение статистики критерия (см. раздел 6.3.4 в [1]).

Целью данной работы является нахождение вычислимых оценок точности аппроксимации статистики Лоули–Хотеллинга нормальным распределением в модели многомерного дисперсионного анализа (MANOVA) для данных большой размерности, когда отношение p/n числа признаков к числу наблюдений стремится к некоторой константе из интервала $(0, 1)$.

В п. 2 сформулирован основной результат работы – теорема 1. Теорема 2 является вспомогательной, но при этом представляет самостоятельный интерес. В теореме 3 приведен неасимптотический результат для статистики Бартлетта–Нанда–Пилая, которая является одной из трех основных статистик, используемых в MANOVA. В п. 4 даны доказательства основных теорем, которые опираются на леммы из п. 3.

Ключевые слова: точность приближений; многомерный дисперсионный анализ; вычислимые оценки; статистика Лоули–Хотеллинга; данные большой размерности.

Исследование финансировалось в рамках государственной поддержки ведущих университетов Российской Федерации “5-100”.

§2. Постановка задачи и основной результат

В рамках многомерного дисперсионного анализа (MANOVA) исследуется следующая многомерная линейная модель: $X = Q\mathbb{B} + \mathcal{E}$, где X – случайная матрица наблюдений размера $N \times p$, Q – неслучайная матрица плана эксперимента размера $N \times k$, $\mathbb{B}$ – неслучайная матрица $k \times p$ регрессионных коэффициентов и $\mathcal{E}$ – матрица ошибок $N \times p$ с распределением $N_{N\times p}(O, I_N \otimes \Sigma)$.

Рассмотрим следующую линейную гипотезу: $H_0 : C\mathbb{B} = O$, где C – известная матрица размера $q \times k$ ранга q. Статистики критериев, инвариантные относительно некоторой группы аффинных преобразований, оказываются функциями от ненулевых собственных значений матрицы $S_h S_e^{-1}$, где

$$S_h = \widehat{\mathbb{B}}^T C^T \left(C \left(Q^T Q\right)^{-1} C^T\right)^{-1} C\widehat{\mathbb{B}} \quad \text{и} \quad S_e = \left(X - Q\widehat{\mathbb{B}}\right)^T \left(X - Q\widehat{\mathbb{B}}\right),$$

при $\widehat{\mathbb{B}} = \left(Q^T Q\right)^{-1} Q^T X$ (см. гл.8 в [2]). Одной из наиболее известных инвариантных статистик, является статистика Лоули–Хотеллинга: $T_0^2 = \operatorname{tr} S_h S_e^{-1}$. В дальнейшем предполагаем, что гипотеза H_0 верна.

В [3] рассмотрен случай большого объёма выборки, т.е. выполнено *условие* **A1:**

$$\mathbf{A1} : p \text{ и } q \text{ фиксированы, } n \to \infty;$$

и получены неасимптотические оценки точности аппроксимации функции распределения статистики Лоули–Хотеллинга:

$$\begin{aligned}\mathbf{P}\{T_0^2 < x\} = G_a(x) + \frac{a}{4n}\{&(q-p-1)\,G_a(x) \\ &- 2qG_{a+2}(x) + (q+p+1)\,G_{a+4}(x)\} + O\left(n^{-2}\right),\end{aligned}$$

где $a = pq$ и G_a – функция распределения хи-квадрат распределения с a степенями свободы и для остаточного члена найдена оценка сверху.

В [4] рассмотрен случай большой размерности данных, т.е. выполнено *условие* **A2**:

$$\mathbf{A2} : q \text{ фиксировано, } p \to \infty, \ n \to \infty, \ \frac{p}{n} \to c \in (0;1);$$

и получено следующее приближение:

$$\mathbf{P}\Big(\frac{1}{\sigma}T_{LH} < z\Big) = \Phi(z) - \phi(z)\Big[\frac{1}{\sqrt{p}}\Big\{\frac{1}{\sigma}b_1 + \frac{1}{\sigma^3}b_3\,H_2\Big(z\Big)\Big\} + \frac{1}{p}\Big\{\frac{1}{\sigma^2}b_2\,H_1\Big(z\Big) + \frac{1}{\sigma^4}b_4\,H_3\Big(z\Big) + \frac{1}{\sigma^6}b_6\,H_5\Big(z\Big)\Big\}\Big] + O\Big(\frac{1}{p\sqrt{p}}\Big),$$

где $T_{LH} = \sqrt{p}\{m\,p^{-1}T_0^2 - q\}$; $\Phi(z)$ есть функция распределения стандартного нормального закона; $m = n - p + q$; $r = p/m$; $\sigma = \sqrt{2q(1+r)}$; $b_i = b_i(r,q)$ суть некоторые функции от r и q; $H_i(z)$ – полиномы Эрмита. При этом результат имел именно асимптотический вид, верхние оценки остаточного члена не находились.

Основными результатами настоящей статьи являются следующие теоремы, дающие оценку точности аппроксимации распределения статистики Лоули–Хотеллинга нормальным распределением для данных большой размерности, т.е. при выполнении условия **A2**:

Теорема 1. *При всех $m > M = M\,(r,q)$ имеет место оценка:*

$$\sup_z \left|\mathbf{P}\Big(\frac{T_{LH}}{\sqrt{2q\,(1+r)}} < z\Big) - \Phi\,(z)\right| \leqslant \frac{K_1\,(r,q)\ln m}{\sqrt{m}},$$

где $K_1\,(r,q)$ – вычислимая функция от r и q.

Отметим, что результат теоремы 1 на логарифмический множитель уступает результату из [4], но превосходит последний в том, что для ошибки погрешности дается вычислимая оценка сверху. При этом само доказательство является новым.

Теорема 2. *Пусть матрицы U и V суть нормированные варианты матриц B и W:*

$$U = (B - pI_q)/\sqrt{p}, \quad V = (W - mI_q)/\sqrt{m}, \tag{1}$$

где B и W независимы и имеют распределения Уишарта $W_q\,(p, I_q)$ и $W_q\,(m, I_q)$ с $m = n - p - q$ соответственно. Если $\operatorname{tr} V^2 < m$, то выполнено следующее неравенство:

$$\begin{aligned}&\left|\sqrt{m}\,\left(\operatorname{tr} BW^{-1} - rq\right) - \left(\sqrt{r}\operatorname{tr} U - r\operatorname{tr} V\right)\right| \\ &\quad\leqslant \frac{\sqrt{r}\,|\operatorname{tr} UV| + \left(rq + \left|\sqrt{r}\operatorname{tr} U - r\operatorname{tr} V\right|/\sqrt{m}\right)\operatorname{tr} V^2}{\sqrt{m} - \operatorname{tr} V^2/\sqrt{m}}.\end{aligned} \tag{2}$$

Заметим, что вероятность события, противоположного событию $\{\operatorname{tr} V^2 < m\}$, фигурирующему в теореме 2, имеет порядок $O\left(1/\sqrt{m}\right)$, как это станет ясно из результатов п. 3.

Напомним определение статистики Бартлетта–Нанда–Пилая, которая является одной из трех оновных статистик, используемых в MANOVA:

$$T_{BNP} = \sqrt{p}\left(1+\frac{p}{m}\right)\left\{\left(1+\frac{m}{p}\right)\operatorname{tr}\left[S_h\left(S_h+S_e\right)^{-1}\right]-q\right\}.$$

Справедлива следующая теорема, дающая оценку точности приближения распределения T_{BNP} нормальным распределением для данных большой размерности, т.е. в условиях **A2**:

Теорема 3. *При всех $m > M = M(r,q)$ справедливо неравенство*

$$\sup_z \left|\mathbf{P}\Big(\frac{T_{BNP}}{\sqrt{2q\,(1+r)}} < z\Big) - \Phi(z)\right| \leqslant \frac{K_2(r,q)\ln m}{\sqrt{m}},$$

где $K_2(r,q)$ – вычислимая функция от r и q.

Доказательству теоремы 3 будет посвящена отдельная статья.

§3. Вспомогательные утверждения

В этой части приведены вспомогательные утверждения, используемые в доказательствах теорем 1 и 2.

Введем дополнительные случайные величины:

$$Z_1 = \operatorname{tr} UV,\ \ Z_2 = \operatorname{tr} V^2,\ \ Z_3 = \operatorname{tr} U - \sqrt{r}\operatorname{tr} V, \tag{3}$$

где случайные матрицы U и V определены в (1).

Положим

$$B = B(q,r,m) = 4\,(q^2+\sqrt{r})(\sqrt{\ln m}+\sqrt{\ln p})^2. \tag{4}$$

Определим также для $i = 1,2,3$ и натуральных m случайные события $A_{i,m}$ как $A_{i,m} = \{\omega : |Z_i(\omega)| \leqslant B\}$.

Положим

$$Z = \frac{\sqrt{r}\,|\operatorname{tr} UV| + \left(rq + |\sqrt{r}\operatorname{tr} U - r\operatorname{tr} V|/\sqrt{m}\right)\operatorname{tr} V^2}{\sqrt{m} - \operatorname{tr} V^2/\sqrt{m}}.$$

Ясно, что существует натуральное $M_1 = M_1(r,q)$, такое что при всех $m \geqslant M_1$ и $\omega \in \cap_{i=1}^{3} A_{i,m}$ имеем

$$Z(\omega) \leqslant \frac{\sqrt{r}\,B + (rq + \sqrt{r}\,B/\sqrt{m})\;B}{\sqrt{m} - B/\sqrt{m}} \leqslant 48\,(2\sqrt{r} + rq)\,(q^2 + \sqrt{r})\frac{\ln m}{\sqrt{m}}. \quad (5)$$

В следующих двух леммах оценим вероятности $\mathbf{P}(A_{i,m}^c)$ для $i = 1,2,3$.

Лемма 1. *Для Z_1 из (3) и B из (4) справедливо неравенство:*

$$\mathbf{P}(|Z_1| > B) \leqslant 6.45\,q^2\,\frac{1 + 1/\sqrt{r}}{\sqrt{m}}. \quad (6)$$

Доказательство леммы 1. Поскольку случайная матрица B имеет распределение Уишарта $W_q\,(p, I_q)$, можем записать $B = X^T X$, где X есть матрица вида

$$X = \begin{pmatrix} X_{11} & \dots & X_{1q} \\ \vdots & \ddots & \vdots \\ X_{p1} & \dots & X_{pq} \end{pmatrix}$$

и $\{X_{ij}\}$ есть совокупность независимых стандартных нормальных случайных величин.

Соответственно, мы имеем следующее представление:

$$B = \left\{ \sum_{k=1}^{p} X_{ki} X_{kj} \right\}_{ij}.$$

Аналогично получаем:

$$W = \left\{ \sum_{l=1}^{m} Y_{li} Y_{lj} \right\}_{ij},$$

где $\{Y_{ij}\}$ есть совокупность независимых стандартных нормальных случайных величин.

Далее, в силу симметричности матриц U и V имеем

$$Z_1 = \operatorname{tr} UV = \sum_{a=1}^{q} (UV)_{aa} = \sum_{a=1}^{q} U_{aa} V_{aa} + \sum_{a \neq b} U_{ab} V_{ab}, \quad (7)$$

где с.в. $U_{aa} = \left(\sum_{k=1}^{p} X_{ka}^2 - p\right)/\sqrt{p}$ имеет нулевое среднее и дисперсию $\mathbf{D}U_{aa} = 2$.

При $a \neq b$ с.в. $U_{ab} = \sum_{k=1}^{p} X_{ka} X_{kb}/\sqrt{p}$ имеет нулевое среднее и дисперсию $\mathbf{D}U_{aa} = 1$.

Аналогичные выкладки верны и для матрицы V.
Из леммы 6 и (7) получаем

$$\begin{aligned}\mathbf{P}(|Z_1| \geqslant B) \leqslant q\big(&\mathbf{P}(|U_{11}| \geqslant \sqrt{B/(2q)}) + \mathbf{P}(|V_{11}| \geqslant \sqrt{B/(2q)})\big)\\ &+ (q^2-q)\mathbf{P}(|U_{12}| \geqslant \sqrt{B/(2(q^2-q))})\\ &+ (q^2-q)\mathbf{P}(|V_{12}| \geqslant \sqrt{B/(2(q^2-q))}).\end{aligned} \tag{8}$$

Для U_{aa} в силу леммы 5 имеем следующее неравенство

$$\sup_x \Big|\mathbf{P}(U_{aa}/\sqrt{2} < x) - \Phi(x)\Big| \leqslant 6.22/\sqrt{p} = 6.22/\sqrt{rm}. \tag{9}$$

Аналогичное неравенство имеет место для V_{aa}:

$$\sup_x \Big|\mathbf{P}(V_{aa}/\sqrt{2} < x) - \Phi(x)\Big| \leqslant 6.22/\sqrt{m}. \tag{10}$$

Оценим теперь U_{ab}, $a \neq b$. По лемме 4 имеем

$$\sup_x |\mathbf{P}(U_{ab} < x) - \Phi(x)| \leqslant C_{BE} \cdot \mathbf{E}\,|X_{1a}X_{1b}|^3/\sqrt{p} \leqslant 2.55/\sqrt{rm}, \tag{11}$$

так как $\mathbf{E}\,|X_{1a}X_{1b}|^3 = 8/\pi$ и константа C_{BE} из неравенства Берри–Эссеена в лемме 4 меньше 1.

Аналогичное неравенство справедливо для V_{ab}:

$$\sup_x |\mathbf{P}(V_{ab} < x) - \Phi(x)| \leqslant 2.55/\sqrt{m}. \tag{12}$$

Отметим, что нетрудно получить (см.,напр., 7.1.13. в [6]) при всех $x \geqslant 0$ неравенство

$$1 - \Phi(x) \leqslant \frac{2}{\pi}\frac{e^{-x^2/2}}{x + \sqrt{x^2 + 8/\pi}}.$$

В частности, при $x \geqslant 1$ имеем

$$1 - \Phi(x) \leqslant 0.23\, e^{-x^2/2}. \tag{13}$$

Объединяя (8)–(13), получаем утверждение леммы 1. □

Лемма 2. *Для Z_2 и Z_3 из* (3) *и B из* (4) *справедливо неравенство:*

$$\mathbf{P}(Z_2 > B) + \mathbf{P}(|Z_3| > B) \leqslant 12.9\, q^2\, \frac{1 + 1/\sqrt{r}}{\sqrt{m}}. \tag{14}$$

Доказательство леммы 2. Аналогично (7) и (8) получаем

$$\begin{aligned}\mathbf{P}(Z_2 \geqslant B) \leqslant\ & q\mathbf{P}(|V_{11}| \geqslant \sqrt{B/(2q)})\\ &+ (q^2-q)\mathbf{P}(|V_{12}| \geqslant \sqrt{B/(2(q^2-q))}).\end{aligned} \tag{15}$$

и

$$\mathbf{P}(|Z_3| \geqslant B) \leqslant q\big(\mathbf{P}(|U_{11}| \geqslant B/(2q))\big) + q\mathbf{P}\big(|V_{11}| \geqslant B/(2q\sqrt{r})\big). \quad (16)$$

Объединяя (9), (10), (13), (15) и (16), получаем утверждение леммы 2. □

Лемма 3. *Пусть случайные величины T, Y и Z определены на одном вероятностном пространстве $(\Omega, \mathbf{A}, \mathbf{P})$, при этом распределение Y является абсолютно непрерывным с ограниченной плотностью $f_Y(z)$. Предположим, что для некоторого события $A \in \mathbf{A}$ при всех $\omega \in A$ выполнено следующее соотношение:*

$$|T(\omega) - Y(\omega)| \leqslant Z(\omega) \leqslant a$$

с некоторой положительной постоянной a. Тогда справедливо неравенство:

$$\sup_x |\mathbf{P}(T < x) - \mathbf{P}(Y < x)| \leqslant \mathbf{P}(A^c) + a \sup_x f_Y(x). \quad (17)$$

Доказательство леммы 3. Заметим, что

$$\sup_x |\mathbf{P}(T < x) - \mathbf{P}(Y < x)| \leqslant \mathbf{P}(A^c) + \sup_x |\mathbf{P}((T < x) \cap A) - \mathbf{P}((Y < x) \cap A)|.$$

Тогда утверждение леммы тривиальным образом следует из соотношений

$$\{T < x\} = \{T - Y + Y < x\} = \{Y < x - (T - Y)\},$$

$$\{Y < x - Z\} \cap A \subset \{Y < x - (T - Y)\} \cap A \subset \{Y < x + Z\} \cap A. \quad \square$$

В следующих двух леммах приводятся два известных результата о скорости сходимости в центральной предельной теореме для независимых одинаково распределённых случайных величин. Первый из результатов относится к случайным величинам без ограничений на тип распределения. Второй результат относится к случайной величине с распределением хи-квадрат, рассматриваемой как сумма независимых одинаково распределённых случайных величин с известным распределением.

Лемма 4. *Пусть случайные величины $\xi_1, \xi_2, \ldots$ независимы и одинаково распределены, выполнено $\mathbf{D}\xi_1 = \sigma^2 > 0$ и существует $\mathbf{E}|\xi_1|^3 < \infty$.*

Тогда для нормированной суммы $T_n = (S_n - \mathbf{E}S_N)/\sqrt{\mathbf{D}S_N}$ *выполнено неравенство:*

$$\sup_x |F_{T_n}(x) - \Phi(x)| \leqslant C_{BE} \frac{\mathbf{E}\,|\xi_1 - \mathbf{E}\xi_1|^3}{\sigma^3\sqrt{n}}$$

с некоторой постоянной C_{BE}.

Известны следующие неравенства (см., напр., [7]) для C_{BE}:

$$\frac{\sqrt{10}+3}{6\sqrt{2\pi}} \leqslant C_{BE} \leqslant 0.4748.$$

Случайная величина с функцией распределения $G_p(x)$, имеющая хи-квадрат распределение с p степенями свободы, может быть представлена в виде суммы p независимых одинаково распределённых случайных величин с хи-квадрат распределением с одной степенью свободы. Этот факт позволяет дать более точные оценки точности аппроксимации нормальным распределением, чем те, которые можно получить в общем случае с помощью неравенства Берри–Эссеена. А именно, имеет место следующий результат (см. лемму 2 в [8]):

Лемма 5. *Для всех* $\lambda \in \left(0; \sqrt{3}-1\right)$ *и целых* $p > 1$ *выполнено*

$$\sup_x \left| G_p\left(p + x\sqrt{2p}\right) - \Phi(x) \right| \leqslant \frac{\min_\lambda \widetilde{D}(\lambda, p)}{\sqrt{p}},$$

где

$$\widetilde{D}\Big(\lambda, p\Big) = \frac{2}{\pi}\left(\frac{\sqrt{\pi}}{6} + \frac{2\Big(1-\lambda\Big)}{\sqrt{p}\Big(2 - 2\lambda - \lambda^2\Big)^2} + \frac{\Big(1+\lambda^2\Big)}{\lambda^2\sqrt{p}}\Big(1+\lambda^2\Big)^{-p/4} + \frac{1}{\lambda^2\sqrt{p}} \exp\Big(-\frac{\lambda^2 p}{4}\Big)\right).$$

Зафиксировав, к примеру, $\lambda = 0.5$, и используя монотонность функции $\widetilde{D}(\lambda, p)$ по параметру p, получим консервативную оценку для $\min_\lambda \widetilde{D}(\lambda, p)$:

$$\min_\lambda \widetilde{D}(\lambda, p) \leqslant \widetilde{D}(0.5, 1) \leqslant 6.22. \tag{18}$$

Лемма 6. *Для любых случайных величин* X *и* Y *и любого действительного числа* $a > 0$ *справедливы неравенства*

$$\mathbf{P}(|X+Y| \geqslant 2a) \leqslant \mathbf{P}(|X| \geqslant a) + \mathbf{P}(|Y| \geqslant a)$$

и

$$\mathbf{P}(|X \cdot Y| \geqslant a^2) \leqslant \mathbf{P}(|X| \geqslant a) + \mathbf{P}(|Y| \geqslant a).$$

Доказательство леммы 6 очевидным образом вытекает из рассуждений от противного. □

Лемма 7. *Если с.в.* $X_1, \ldots, X_k$ *независимы и таковы, что* $|\mathbf{P}(X_j \leqslant x) - \Phi(x)| \leqslant D_j$ *при всех* x *и* $j = 1, \ldots, k$ *с некоторыми постоянными* $D_1, \ldots, D_k$*, то*

$$\Big|\mathbf{P}\Big(\sum_{j=1}^{k} c_j\, X_j \leqslant x\Big) - \Phi(x)\Big| \leqslant \sum_{j=1}^{k} D_j,$$

где $c_1, \ldots, c_k$ *суть произвольные постоянные, для которых* $c_1^2 + \cdots + c_k^2 = 1$.

Доказательство леммы 7 см., например, в теореме 3.1 в [9]]. □

§4. Доказательства теорем 1–2

Начнем с доказательства теоремы 2, поскольку неравенство (2) является ключевым в доказательстве теоремы 1.

Доказательство теоремы 2. Воспользовавшись матричным равенством

$$(I + A)^{-1} - (I - A) = A^2\,(I + A)^{-1},$$

из определения (1) получаем

$$\begin{aligned}
&\sqrt{m}\Big(BW^{-1} - rI_q\Big) - \Big(\sqrt{r}U - rV\Big) \\
&\quad = -\sqrt{\frac{r}{m}}UV + \sqrt{m}\Big(rI_q + \frac{1}{\sqrt{m}}\sqrt{r}U\Big)\Big(\Big(I_q + \frac{1}{\sqrt{m}}V\Big)^{-1} - \Big(I_q - \frac{1}{\sqrt{m}}V\Big)\Big) \\
&\quad = -\sqrt{\frac{r}{m}}UV + \frac{1}{\sqrt{m}}\Big(rI_q + \frac{1}{\sqrt{m}}\sqrt{r}U\Big)V^2\Big(I_q + \frac{1}{\sqrt{m}}V\Big)^{-1}.
\end{aligned}$$

Отсюда для следов этих матриц имеем следующее неравенство:

$$
\begin{aligned}
&\left|\sqrt{m}\left(\operatorname{tr} BW^{-1} - rq\right) - \left(\sqrt{r}\operatorname{tr} U - r\operatorname{tr} V\right)\right| \\
&\leqslant \frac{1}{\sqrt{m}}\sqrt{r}\left|\operatorname{tr} UV\right| + \frac{1}{\sqrt{m}}\left|\operatorname{tr}\left[\left(rI_q + \frac{1}{\sqrt{m}}\sqrt{r}U\right)V^2\left(I_q + \frac{1}{\sqrt{m}}V\right)^{-1}\right]\right| \\
&\leqslant \frac{1}{\sqrt{m}}\sqrt{r}\left|\operatorname{tr} UV\right| + \frac{1}{\sqrt{m}}\operatorname{tr} V^2 \operatorname{tr}\left[\left(rI_q + \frac{1}{\sqrt{m}}\sqrt{r}U\right)\left(I_q + \frac{1}{\sqrt{m}}V\right)^{-1}\right] \\
&= \frac{1}{\sqrt{m}}\sqrt{r}\left|\operatorname{tr} UV\right| + \frac{1}{\sqrt{m}}\operatorname{tr} V^2 \operatorname{tr} BW^{-1}. \qquad (19)
\end{aligned}
$$

Для получения предпоследнего неравенства мы использовали симметричность и неотрицательную определённость обеих случайных матриц V^2 и

$$
\left[\left(rI_q + \frac{1}{\sqrt{m}}\sqrt{r}U\right)\left(I_q + \frac{1}{\sqrt{m}}V\right)^{-1}\right] = BW^{-1},
$$

поскольку для симметричных неотрицательно определённых матриц X и Y выполнено соотношение (см. [5]): $\operatorname{tr} XY \leqslant \operatorname{tr} X \operatorname{tr} Y$.

Мы видим, что случайная величина $\operatorname{tr} BW^{-1}$ фигурирует в крайней левой и крайней правой частях неравенства (19). Преобразуя полученное неравенство, получаем, что при $\operatorname{tr} V^2 < m$ выполнено (2). Тем самым доказательство теоремы 2 завершено. □

Переходим к доказательству теоремы 1.

Доказательство теоремы 1. Используя лемму 1 из [4], перейдём к представлению статистики Лоули–Хотеллинга

$$
T_{LH} = \sqrt{p}\{m\, p^{-1} \operatorname{tr} S_h S_e^{-1} - q\}
$$

в терминах матриц B и W размера $q \times q$ вместо матриц S_h и S_e размера $p \times p$, где S_h и S_e определены в (1), а матрицы B и W независимы и имеют распределения Уишарта $W_q\,(p, I_q)$ и $W_q\,(m, I_q)$ с $m = n - p - q$ соответственно. При этом будем пользоваться следующим соотношением (см. [4]): $\operatorname{tr} S_h S_e^{-1} = \operatorname{tr} BW^{-1}$.

Согласно (2) для Z_1, Z_2 и Z_3 (см. определение в (3)) при $Z_2 < m$ имеем

$$
\begin{aligned}
\left|\sqrt{r}\, T_{LH} - \left(\sqrt{r}\operatorname{tr} U - r\operatorname{tr} V\right)\right| &= \left|\sqrt{m}\Big(\operatorname{tr} BW^{-1} - rq\Big) - \Big(\sqrt{r}\operatorname{tr} U - r\operatorname{tr} V\Big)\right| \\
&\leqslant \frac{\sqrt{r}\Big|Z_1\Big| + \Big(rq + \sqrt{r}\Big|Z_3\Big|/\sqrt{m}\Big)Z_2}{\sqrt{m} - Z_2/\sqrt{m}}.
\end{aligned}
$$

Следовательно, в силу (5) и (17) при всех $m \geqslant M_1$ получаем

$$\sup_z \left| \mathbf{P}\left(\frac{T_{LH}}{\sqrt{2q\Big(1+r\Big)}} < z \right) - \mathbf{P}\left(\frac{\operatorname{tr} U - \sqrt{r}\operatorname{tr} V}{\sqrt{2q\Big(1+r\Big)}} < z \right) \right|$$

$$\leqslant \sum_{i=1}^{3} \mathbf{P}(|Z_i| > B) + K_3(r,q)\,\frac{\ln m}{\sqrt{m}} \cdot \sup_x f(x), \tag{20}$$

где $f(x)$ есть плотность с.в. $(\operatorname{tr} U - \sqrt{r}\operatorname{tr} V)/\sqrt{2q\,(1+r)}$ и $K_3\,(r,q)$ – некоторая вычислимая функция от r и q.

Отметим, что, поскольку матрицы B и W независимы, матрицы U и V также независимы между собой. Известно (см., напр., гл.2 в [1]), что $\operatorname{tr} B$ и $\operatorname{tr} W$ имеют хи-квадрат распределения с pq и mq степенями свободы соответственно. Известно также, что плотность распределения хи-квадрат с $k \geqslant 3$ степенями свободы ограничена сверху величиной: $1/(2\sqrt{\pi\,(k-2)})$. Поэтому для плотности $f(x)$ справедлива равномерная оценка

$$f(x) \leqslant \min\left(\frac{\sqrt{p}}{\sqrt{(pq-2)}},\ \frac{\sqrt{m}}{\sqrt{r\,(qm-2)}} \right) \cdot \frac{\sqrt{q(1+r)}}{\sqrt{2\pi}}. \tag{21}$$

Объединяя утверждения лемм 5 и 7 и соотношения (1), (6), (14), (18), (20) и (21), получаем утверждение теоремы 1. □

Список литературы

1. Y. Fujikoshi, V. V. Ulyanov, R. Shimizu, *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. Wiley Series in Probability and Statistics, Wiley, Hoboken, N.J., 2010.
2. T. W. Anderson, *An Introduction to Multivariate Analysis*. 3rd ed., Wiley, NewYork, 2003.
3. Y. Fujikoshi, V. V. Ulyanov, R. Shimizu, *L_1-norm error bounds for asymptotic expansions of multivariate scale mixtures and their applications to Hotelling's generalized T_0^2*. — J. Multivariate Anal. **96**, No. 1 (2005), 1–19.
4. H. Wakaki, Y. Fujikoshi, V. V. Ulyanov, *Asymptotic expansions of the distributions of MANOVA test statistics when the dimension is large.* — Hiroshima Math. J. **44**, No. 3 (2014), 247–259.
5. I. D. Coope, *On matrix trace inequalities and related topics for products of Hermitian matrices*. — J. Math. Anal. Appl. **188**, No. 3 (1949), 999–1001.
6. M. Abramowitz, I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards Applied Mathematics Series, 55, U.S. Government Printing Office, Washington, D.C., 1964.

7. I. G. Shevtsova, *On the absolute constants in the Berry-–Esseen type inequalities for identically distributed summands*, (2011) `Available at: http://arxiv.org/pdf/1111.6554v1.pdf.`
8. Ю. Кавагучи, В. В. Ульянов, Я. Фуджикоши, *Приближения для статистик, описывающих геометрические свойства данных большой размерности, с оценками ошибок.* — Информатика и её примен. **4**, Вып. 1 (2010), 22—27.
9. V. V. Ulyanov, H. Wakaki, Y. Fujikoshi, *Berry-Esseen bound for high dimensional asymptotic approximation of Wilks' Lambda distribution.* — Statist. and Probab. Letters **76**, No. 12 (2006), 1191—1200.

Lipatiev A. A. Ulyanov V. V. Non-asymptotic analysis of Lawley–Hotelling Statistic for high dimensional data.

We consider General Linear Model (GLM) that includes multivariate analysis of variance (MANOVA) and multiple linear regression as special cases. In practice, there are several widely used criteria for GLM: Wilks' lambda, Bartlett–Nanda–Pillai test, Lawley–Hotelling test and Roy maximum root test. Limiting distributions for the first three mentioned tests are known under different asymptotic settings. In the present paper we get the computable error bounds for normal approximation of Lawley–Hotelling statistic when dimensionality grows proportionally to sample size. This result enables us to get more precise calculations of the p-values in applications of multivariate analysis. In practice, more and more often analysts encounter situations when the number of factors is large and comparable with the sample size. Examples include medicine, biology (i.e., DNA microarray studies) and finance.

Московский государственный
университет им. М. В. Ломоносова,
ГСП-1 Москва, 119991 Россия
E-mail: `allipatev@cs.msu.ru`

Московский государственный
университет им. М. В. Ломоносова,
ГСП-1 Москва, 119991 Россия;
Национальный исследовательский университет
Высшая школа экономики
ул. Мясницкая, д. 20,
Москва, 101000 Россия
E-mail: `vulyanov@cs.msu.ru`

Поступило 5 ноября 2019 г.