ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2010. Т. 4. Вып. 1. С. 22–27

ПРИБЛИЖЕНИЯ ДЛЯ СТАТИСТИК, ОПИСЫВАЮЩИХ ГЕОМЕТРИЧЕСКИЕ СВОЙСТВА ДАННЫХ БОЛЬШОЙ РАЗМЕРНОСТИ, С ОЦЕНКАМИ ОШИБОК*

Ю. Кавагучи[1], В. В. Ульянов[2], Я. Фуджикоши[3]

Аннотация: При описании геометрических свойств n наблюдений с p признаками необходимо исследовать асимптотическое поведение трех статистик: длины p-мерного вектора наблюдений, расстояния между двумя векторами наблюдений и угла между ними. В [1] найдено асимптотическое поведение указанных статистик, когда размерность p стремится к бесконечности, а объем выборки n остается постоянным. В настоящей работе результаты [1] уточняются путем построения асимптотических разложений. Кроме этого, найдены вычислимые оценки погрешности приближения предельными распределениями для первых двух статистик. Эти результаты будут также полезны в ситуациях, когда размерность данных не является чрезмерно большой.

Ключевые слова: данные большой размерности; асимптотические разложения; оценки погрешности; геометрические свойства

1 Введение

Рассмотрим случайную выборку $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, взятую из p-мерного распределения. Набор данных можно рассматривать как n векторов или точек в p-мерном пространстве. При анализе данных полезным оказывается исследование следующих трех геометрических характеристик: длины p-мерного вектора наблюдений, расстояния между двумя векторами наблюдений и угла между ними.

В последние годы значительный интерес вызывают исследования для данных большой размерности. Это связано с тем, что такого типа данные встречаются все в большем числе приложений, в частности при анализе финансовых рынков и деятельности финансовых организаций, в социологии, генетике, биологии и математической физике. В асимптотической теории для данных большой размерности предполагается, что либо (i) оба параметра — размерность p и объем выборки n — стремятся к бесконечности, либо (ii) p стремится к бесконечности, а объем n фиксирован.

Недавний результат в рамках предположения (i) представлен, например, в [2], где получены также вычислимые оценки точности приближений.

Если $\boldsymbol{x}_i$ взяты из нормального распределения $N(0, I_p)$, в [1] в предположении (ii) показано, что три статистики, описывающие геометрические свойства данных, удовлетворяют следующим предельным соотношениям:

$$\|\boldsymbol{x}_i\| = \sqrt{p} + \mathrm{O}_p(1), \qquad i = 1, \ldots, n\,;$$
$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\| = \sqrt{2p} + \mathrm{O}_p(1), \qquad i, j = 1, \ldots n\,,\ \ i \neq j\,;$$
$$\mathrm{ang}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{2}\pi + \mathrm{O}_p(p^{-1/2}),\ \ i, j = 1, \ldots n,\ i \neq j\,,$$

где $\|\cdot\|$ — евклидово расстояние и O_p обозначает стохастический порядок малости. Из этих результатов вытекает, что при увеличении размерности данные сходятся к вершинам правильного симплекса.

В настоящей работе эти результаты сначала уточняются путем построения асимптотических разложений для распределений указанных трех статистик. Затем строятся вычислимые оценки погрешности для предельных распределений первых двух статистик. При этом получены два типа неравенств с использованием идей из работы [3], где найдена оценка порядка $\mathrm{O}(p^{-1})$ хи-квадрат приближения для преобразованной хи-квадрат случайной величины с p степенями свободы. Тем самым получена возможность судить о геометрических свойствах статистических данных, размерность которых не является чрезмерно большой.

*Исследования выполнены при частичной поддержке РФФИ, гранты 08-01-00567, 08-01-91205, 09-01-12180.

[1]Высшая научная и инженерная школа, Университет Чуо, Токио, n15007@gug.math.chuo-u.ac.jp

[2]Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, vulyan@gmail.com

[3]Высшая научная и инженерная школа, Университет Чуо, Токио, fuji@math.sci.hiroshima-u.ac.jp

2 Асимптотические разложения для распределений статистик

Пусть $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ $(i = 1, \ldots, n)$ есть выборка из распределения $N(0, I_p)$.

Рассмотрим сначала статистику

$$\|\boldsymbol{x}_i\| = S_{ii}^{1/2}\,,$$

где $S_{ii} = \sum\limits_{k=1}^{p} x_{ik}^2$. Поскольку статистика S_{ii} распределена как χ_p^2 — хи-квадрат случайная величина с p степенями свободы, она не зависит от i. Положим

$$V = \frac{S_{ii} - p}{\sqrt{2p}}\,.$$

Распределение V сходится к стандартному нормальному при стремлении p к бесконечности. Поскольку $S_{ii} = p(1 + \sqrt{2}\,p^{-1/2}V)$, имеем

$$\|\boldsymbol{x}_i\| = S_{ii}^{1/2} = \sqrt{p}\left(1 + \frac{1}{2}\sqrt{2}Vp^{-1/2} - \right.$$
$$\left. - \frac{1}{4}Vp^{-1} + \frac{1}{8}\sqrt{2}V^3 + \cdots\right)\,.$$

Рассмотрим асимптотическое разложение для распределения случайной величины

$$T_1 = \sqrt{2}(\|\boldsymbol{x}_i\| - \sqrt{p})\,.$$

Характеристическая функция величины T_1 запишется в виде

$$C_{T_1}(t) = \mathrm{E}\left[\exp\left\{(it)\sqrt{2}\,(\|\boldsymbol{x}_i\| - \sqrt{p})\right\}\right] =$$
$$= C_0(t) + \frac{1}{\sqrt{p}}\,C_1(t) + \frac{1}{p}\,C_2(t) + \mathrm{O}\left(p^{-3/2}\right)\,,$$

где

$$C_0(t) = \mathrm{E}\left[\exp\{(it)V\}\right]\,;$$
$$C_1(t) = \mathrm{E}\left[-\frac{1}{4}\sqrt{2}(it)V^2\exp\{(it)V\}\right]\,;$$
$$C_2(t) = \mathrm{E}\left[\left(\frac{1}{4}\,(it)V^3 + \frac{1}{16}\,(it)^2V^4\right)\exp\{(it)V\}\right]\,.$$

Для вычисления функций $C_0(t)$, $C_1(t)$ и $C_2(t)$ используем асимптотическое разложение для плотности случайной величины (с.в.) V. Характеристическая функция с.в. V может быть записана в виде

$$C_V(t) = \exp\left\{\frac{1}{2}\,(it)^2\right\}\left[1 + \frac{1}{\sqrt{p}}\,\frac{1}{6}(it)^3\kappa_3 + \right.$$
$$\left. + \frac{1}{p}\left\{\frac{1}{72}\,(it)^6\kappa_3^2 + \frac{1}{24}\,(it)^4\kappa_4\right\} + \mathrm{O}\left(p^{-3/2}\right)\right]\,.$$

Соответственно, плотность с.в. V раскладывается как

$$\phi(x)\left[1 + \frac{1}{\sqrt{p}}\,\frac{\sqrt{2}}{3}\,h_3(x) + \right.$$
$$\left. + \frac{1}{p}\left\{\frac{1}{9}\,h_6(x) + \frac{1}{2}\,h_4(x)\right\}\right] + \mathrm{O}\left(p^{-3/2}\right)\,,$$

где $h_i(x)$ — многочлен Эрмита порядка i, а $\phi(x)$ — плотность стандартного нормального распределения. Используя это разложение, можно найти $C_0(t)$, $C_1(t)$ и $C_2(t)$. Асимптотическое разложение для распределения с.в. T_1 можно получить, обращая ее характеристическую функцию. Тем самым доказана следующая теорема:

Теорема 1. *Пусть $\boldsymbol{x}_i$ есть p-мерный случайный вектор с распределением $N(0, I_p)$. Тогда функция распределения с.в. $T_1 = \sqrt{2}(\|\boldsymbol{x}_i\| - \sqrt{p})$ может быть представлена в виде*

$$\Phi(x) - \phi(x)\left[\frac{1}{\sqrt{p}}\,\ell_1(x) + \frac{1}{p}\,\ell_2(x)\right] + \mathrm{O}(p^{-3/2})\,,$$

где $\Phi(x)$ — функция распределения стандартного нормального закона, а $\ell_1(x)$ и $\ell_2(x)$ определяются формулами

$$\ell_1(x) = \frac{1}{12}\,\sqrt{2}h_2(x) - \frac{1}{4}\,\sqrt{2}h_0(x)\,;$$
$$\ell_2(x) = \frac{1}{144}\,\left[-15h_5(x) - 6h_3(x) + 16h_2(x) - \right.$$
$$\left. - 81h_1(x) + 72h_0(x)\right]\,.$$

Распределение с.в. $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ $(i \neq j)$ тесно связано с распределением $\|\boldsymbol{x}_i\|$, так как $(x_{ik} - x_{jk})/\sqrt{2} \sim$ $\sim N(0, 1)$. Точнее, распределение с.в. $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|/\sqrt{2}$ совпадает с распределением с.в. $\|\boldsymbol{x}_i\|$. Следовательно, распределение T_1 совпадает с распределением

$$T_2 = \sqrt{2}\left\{\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\sqrt{2}} - \sqrt{p}\right\} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\| - \sqrt{2p}$$

и справедливо следствие:

Следствие 1. *Пусть $\boldsymbol{x}_i$ и $\boldsymbol{x}_j$ — независимые случайные векторы с распределением $N(0, I_p)$. Тогда распределение $T_2 = \|\boldsymbol{x}_i - \boldsymbol{x}_j\| - \sqrt{2p}$ совпадает с распределением T_1. В частности, для функции распределения T_2 имеет место то же асимптотическое разложение, что и для T_1.*

Далее рассмотрим асимптотическое разложение для распределения угла θ между двумя независимыми векторами $\boldsymbol{x}_i$ и $\boldsymbol{x}_j$, взятыми из нормального распределения $N(0, I_p)$. Поскольку

$$\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 = S_{ii} + S_{jj} - 2S_{ij}\,,$$

где $S_{ij} = \sum\limits_{k=1}^{p} x_{ik}x_{jk}$, получаем

$$\cos\theta = \frac{\|\boldsymbol{x}_i\|^2 + \|\boldsymbol{x}_j\|^2 - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|} = \\ = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} = r_{ij}\,,$$

где r_{ij} — выборочный коэффициент корреляции. Распределение r_{ij} исследовалось во многих работах. В частности, если $\rho_{ij} = 0$, для функции распределения r_{ij} имеем [4]:

$$\mathrm{P}\left(\sqrt{p}\cdot r_{ij} < x\right) = \\ = \Phi(x) + \frac{1}{p}\left\{-\frac{3}{4}\,x + \frac{1}{4}\,x^3\right\}\phi(x) + \mathrm{O}\left(p^{-3/2}\right)\,.$$

Поскольку θ может быть разложен в терминах r_{ij} в виде

$$\theta = \arccos r_{ij} = \frac{1}{2}\,\pi - \left(r_{ij} + \frac{1}{6}\,r_{ij}^3 + \frac{3}{40}\,r_{ij}^5 + \cdots\right)\,,$$

полагая $y = \sqrt{p}\cdot r_{ij}$, имеем

$$\sqrt{p}\left(\frac{1}{2}\,\pi - \theta\right) = y + \frac{1}{6p}\,y^3 + \mathrm{o}\left(p^{-1}\right)\,.$$

Тем самым характеристическую функцию $T_3 = \sqrt{p}\,(\pi/2 - \theta)$ можно записать в виде

$$C_{\mathrm{ang}}(it) = C_0(t) + \frac{1}{6p}\,C_1(t) + \mathrm{o}(p^{-1})\,,$$

где

$$C_0(t) = \mathrm{E}\left[\exp\left\{(it)y\right\}\right]\,;$$
$$C_1(t) = \mathrm{E}\left[\exp\left\{(it)y\right\}(it)y^3\right]\,.$$

Используя асимптотическое разложение для распределения $\sqrt{p}\cdot r_{ij}$, вычисляем $C_0(t)$ и $C_1(t)$. Обращая полученную характеристическую функцию, приходим к следующей теореме:

Теорема 2. *Пусть $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ есть p-мерный случайный вектор, взятый из нормального распределения $N(0, I_p)$, и θ — угол между двумя независимыми векторами $\boldsymbol{x}_i$ и $\boldsymbol{x}_j$. Тогда функция распределения с.в. $T_3 = \sqrt{p}\,(\pi/2 - \theta)$ может быть представлена в виде*

$$\Phi(x) + \frac{1}{12p}\,\left[h_3(u) - 6h_1(x)\right]\phi(x) + \mathrm{o}(p^{-1})\,.$$

3 Оценки погрешностей приближений

В этой части получены вычислимые выражения $B(p)$ такие, что

$$|\mathrm{P}(T_1 \le x) - \Phi(x)| \le B(p) = \mathrm{O}(p^{-1/2})\,, \qquad (1)$$

где $T_1 = \sqrt{2}(\|\boldsymbol{x}_i\| - \sqrt{p})$. В силу следствия 1 результат (1) справедлив и для $T_2 = \|\boldsymbol{x}_i - \boldsymbol{x}_j\| - \sqrt{2p}$.

Идея доказательства опирается на работу [3]. Рассмотрим случайную величину

$$V_p = \frac{\chi_p^2 - p}{\sqrt{2p}}\,.$$

Пусть $h(x)$ есть действительная функция, определяемая формулой

$$h(x) = \sqrt{2}\left[\left(\sqrt{2p}\cdot x + p\right)^{1/2} - \sqrt{p}\right]\,.$$

Тогда

$$h(V_p) = \sqrt{2}\left(\sqrt{\chi_p^2} - \sqrt{p}\right) = \sqrt{2}\,(\|X\| - \sqrt{p}) = T_1\,.$$

Если $x \ge -\sqrt{2p}$, то

$$\sup_{-\sqrt{2p}\,\le\,x} |\mathrm{P}(T_1 \le x) - \Phi(x)| = \\ = \sup_{-\sqrt{2p}\,\le\,x} |\mathrm{P}(h(V_p) \le x) - \Phi(x)| = \\ = \sup_{-\sqrt{2p}\,\le x}\left|\mathrm{P}\left(V_p \le\, x + \frac{x^2}{2\sqrt{2p}}\right) - \Phi(x)\right| \le \\ \le \sup_{-\sqrt{2p}\,\le\,x}\left(|I_1| + |I_2|\right)\,,$$

где

$$I_1 = \mathrm{P}\left(V_p \le x + \frac{x^2}{2\sqrt{2p}}\right) - \Phi\left(x + \frac{x^2}{2\sqrt{2p}}\right)\,;$$
$$I_2 = \Phi\left(x + \frac{x^2}{2\sqrt{2p}}\right) - \Phi(x)\,.$$

Для I_1 в следующем разделе (см. леммы 1 и 2 ниже) будут получены два типа оценок:

$$|I_1| = \left|\mathrm{P}\left(V_p \le x + \frac{x^2}{2\sqrt{2p}}\right) - \Phi\left(x + \frac{x^2}{2\sqrt{2p}}\right)\right| \le \\ \le D_i(\lambda, p) \text{ для } i = 1, 2.$$

Здесь же займемся оцениванием I_2:

$$|I_2| = \left|\Phi\left(V_p \le x + \frac{x^2}{2\sqrt{2p}}\right) - \Phi(x)\right| = \\ = \left|\frac{1}{\sqrt{2\pi}}\int\limits_{x}^{x+x^2/(2\sqrt{2p})} \exp\left\{-\frac{z^2}{2}\right\}\,dz\right|\,.$$

Если $x \geq 0$, то для I_2 получаем

$$I_2 \leq \frac{x^2}{2\sqrt{2p}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \leq \frac{1}{2e\sqrt{p\pi}}\,.$$

Если $-\sqrt{2p} \leq x \leq 0$, то $x + x^2/(2\sqrt{2p}) \leq x/2$. Следовательно,

$$I_2 \leq \frac{x^2}{2\sqrt{2p}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(x + \frac{x^2}{2\sqrt{2p}}\right)^2\right\} \leq$$
$$\leq \frac{x^2}{2\sqrt{2p}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x}{2}\right)^2\right\} \leq \frac{2}{e\sqrt{p\pi}}\,.$$

Поскольку $1/(2e\sqrt{p\pi}) < 2/(e\sqrt{p\pi})$, при выполнении условия $x \geq -\sqrt{2p}$ имеем

$$\sup_{-\sqrt{2p} \leq x} |\mathrm{P}(T_1 \leq x) - \Phi(x)| \leq D(\lambda, p) + \frac{2}{e\sqrt{p\pi}}\,, \quad (2)$$

где $D(\lambda, p) = \min\left(D_1(\lambda, p), D_2(\lambda, p)\right)$.

Известно (см., например, 7.1.13 в [5]), что для любого $x \geq 0$

$$e^{x^2} \int\limits_x^\infty e^{-t^2}\, dt \leq \frac{1}{x + \sqrt{x^2 + 4/\pi}}\,. \quad (3)$$

Заменой переменных из (3) для всех $x > 0$ получаем

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

Тогда

$$\Phi(-\sqrt{2p}) = 1 - \Phi(\sqrt{2p}) \leq$$
$$\leq \sqrt{\frac{2}{\pi}} \frac{e^{-p}}{\sqrt{2p} + \sqrt{2p + 8/\pi}}\,. \quad (4)$$

Из (2) и (4) получаем следующую оценку

$$\sup_{x \in R} |\mathrm{P}(T_1 \leq x) - \Phi(x)| \leq$$
$$\leq \max\left(\min_\lambda D(\lambda, p) + \frac{2}{e\sqrt{p\pi}},\right.$$
$$\left.\sqrt{\frac{2}{\pi}} \frac{e^{-p}}{\sqrt{2p} + \sqrt{2p + 8/\pi}}\right)\,.$$

Поскольку

$$\frac{2}{e\sqrt{p\pi}} > \sqrt{\frac{2}{\pi}} \frac{e^{-p}}{\sqrt{2p} + \sqrt{2p + 8/\pi}}$$

и

$$\min_\lambda D(\lambda, p) > 0\,,$$

имеем

$$\sup_{x \in R} |\mathrm{P}(T_1 \leq x) - \Phi(x)| \leq \min_\lambda D(\lambda, p) + \frac{2}{e\sqrt{p\pi}}\,.$$

Таким образом (см. (6)), доказана следующая теорема:

Теорема 3. *Справедлива оценка*

$$\sup_{x \in R} |\mathrm{P}(T_1 \leq x) - \Phi(x)| < B(p)\,,$$

где

$$B(p) = \min_\lambda D(\lambda, p) + \frac{2}{e\sqrt{p\pi}}\,, \quad (5)$$

минимум берется по всем $\lambda \in (0, \sqrt{3} - 1)$ *и*

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

Из следствия 1 и оценки для T_1 также получается

Следствие 2. *Для статистики* $T_2 = \|\boldsymbol{x}_i - \boldsymbol{x}_j\| - \sqrt{2p}$

$$\sup_{x \in R} |\mathrm{P}(T_2 \leq x) - \Phi(x)| < B(p)\,,$$

где $B(p)$ определено в (5).

4 Два подхода к оценке I_1

Будем использовать результат из [3] для нахождения величины $D(\lambda, p)$, которая будет выступать оценкой для I_1. В [3] доказано, что

$$\sup_{x \in R} |F_p(x) - \Phi_p(x)| \leq D_0(\lambda, p)\,,$$

где

$$F_p(x) = \mathrm{P}(V_p \leq x) = \mathrm{P}\left(\chi_p^2 - p \leq \sqrt{2p}x\right)\,;$$
$$\Phi_p(x) = \Phi(x) + \frac{\sqrt{2}(1 - x^2)}{3\sqrt{p}}\,\phi(x)$$

и

$$D_0(\lambda, p) = \frac{2}{\pi p}\left(\frac{4}{9} + \frac{2(1-\lambda)}{(2 - 2\lambda - \lambda^2)^2} + \right.$$
$$\left. + \frac{(1+\lambda^2)^{1-p/4}}{\lambda^2} + \frac{3+\lambda}{3\lambda^2} e^{-\lambda^2(3-\lambda)p/(12+4\lambda)}\right)\,.$$

Чтобы оценить I_1, надо найти равномерную оценку для $F_p(x) - \Phi(x)$. С этой целью будем использовать два подхода. Первый подход очень прост. Используя результат из [3]

$$\sup_{x\in\mathrm{R}} |F_p(x) - \Phi(x)| =$$
$$= \sup_{x\in\mathrm{R}} \left| F_p(x) - \left(\Phi_p(x) - \frac{\sqrt{2}(1-x^2)}{3\sqrt{p}} \phi(x) \right) \right| \le$$
$$\le \sup_{x\in\mathrm{R}} |F_p(x) - \Phi_p(x)| + \frac{\sqrt{2}}{3\sqrt{p}} \sup_{x\in\mathrm{R}} \left|(1-x^2)\phi(x)\right| \le$$
$$\le D_0(\lambda, p) + \frac{1}{3\sqrt{p\pi}}\,,$$

получаем лемму:

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

$$\sup_{x\in\mathrm{R}} |F_p(x) - \Phi(x)| \le D_1(\lambda, p)\,,$$

где

$$D_1(\lambda, p) = \frac{2}{\pi p} \left(\frac{4}{9} + \frac{2(1-\lambda)}{(2-2\lambda-\lambda^2)^2} + \right.$$
$$+ \frac{(1+\lambda^2)^{1-p/4}}{\lambda^2} +$$
$$\left. + \frac{3+\lambda}{3\lambda^2} e^{-\lambda^2(3-\lambda)p/(12+4\lambda)} \right) + \frac{1}{3\sqrt{p\pi}}\,.$$

Второй подход основан на модификации результата из [3]. Пусть $f_p(t)$ и $g(t)$ — характеристические функции распределений $F_p(x)$ и $\Phi(x)$ соответственно. Положим $p^* = p/2$. Имеем

$$f_p(t) = \exp(-it\sqrt{p^*}) \left(1 - \frac{it}{\sqrt{p^*}}\right)^{-p^*}; \;\; g(t) = e^{-t^2/2}\,.$$

Из формулы обращения для характеристических функций получаем

$$|F_p(x) - \Phi(x)| \le \frac{1}{2\pi}(I_1 + I_2 + I_3)\,,$$

где

$$I_1 = \int\limits_{|t|\,<\,\lambda\sqrt{p^*}} \frac{1}{|t|} \left| f_p(t) - e^{-t^2/2} \right| dt\,;$$
$$I_2 = \int\limits_{|t|\,\ge\,\lambda\sqrt{p^*}} \frac{|f_p(t)|}{|t|}\, dt\,;$$
$$I_3 = \int\limits_{|t|\,\ge\,\lambda\sqrt{p^*}} \frac{e^{-t^2/2}}{|t|}\, dt\,.$$

Используя ту часть комплексной функции $\tau(z) = \log(1-z)$, для которой $\tau(0) = 0$, запишем:

$$f_p(t) = \exp\left\{ -it\sqrt{p^*} - p^* \log\left(1 - \frac{it}{\sqrt{p^*}}\right) \right\} =$$
$$= \exp\left\{ -\frac{t^2}{2} - \frac{it^3}{3\sqrt{p^*}} + p^* R_{p^*}(t) \right\} =$$
$$= \exp\left\{ -\frac{t^2}{2} - \frac{it^3}{3\sqrt{p^*}} \right\} + S_{p^*}(t)\,,$$

где

$$R_{p^*}(t) = -\log\left(1 - \frac{it}{\sqrt{p^*}}\right) - \frac{it}{\sqrt{p^*}} - \frac{(it)^2}{2p^*} - \frac{(it)^3}{3{p^*}^{3/2}}\,;$$
$$S_{p^*}(t) = \exp\left\{ -\frac{t^2}{2} - \frac{it^3}{3\sqrt{p^*}} \right\} (\exp p^* R_{p^*}(t) - 1)\,.$$

В [3] для $|t| < \lambda\sqrt{p^*}$ показано, что

$$|p^* R_{p^*}(t)| \le \frac{|t|^4}{4p^* \left(1 - |t|/\sqrt{p^*}\right)}$$

и

$$|S_{p^*}(t)| \le e^{-t^2/2} \left|\exp\{p^* R_{p^*}(t)\} - 1\right| \le$$
$$\le e^{-t^2/2} \frac{|t|^4}{4p^*(1-\lambda)} \exp\left\{ \frac{|t|^2\lambda^2}{4p^*(1-\lambda)} \right\} =$$
$$= \frac{|t|^4}{4p^*(1-\lambda)} e^{-(-a|t|^2)/2}$$

с $a = 1 - \lambda^2/(2(1-\lambda))$. Следовательно,

$$I_1 \le I_{11} + I_{12}\,,$$

где

$$I_{11} = \int\limits_{|t|\,<\,\lambda\sqrt{p^*}} \frac{e^{-t^2/2}}{|t|} \left| e^{-it^3/(3\sqrt{p^*})} - 1 \right| dt\,;$$
$$I_{12} = \int\limits_{|t|\,<\,\lambda\sqrt{p^*}} \frac{|S_{p^*}|}{|t|}\, dt\,.$$

Используя неравенство $|e^{-iz} - 1| \le |z|$, получаем

$$I_{11} \le \sqrt{\frac{\pi}{2}}\, \frac{2}{3\sqrt{p^*}}$$

и

$$I_{12} \le \frac{1}{4p^*(1-\lambda)} \int\limits_{|t|\,<\,\lambda\sqrt{p^*}} |t|^3 e^{-at^2/2}\, dt \le$$
$$\le \frac{2}{4p^*(1-\lambda)} \int\limits_0^\infty t^3 e^{-at^2/2}\, dt =$$
$$= \frac{4}{4p^*(1-\lambda)a^2} = \frac{2}{p(1-\lambda)a^2}\,,$$

где $a = 1 - \lambda^2/(2-2\lambda)$. Заметим, что $|f_p(t)| = (1+t^2/p^*)^{-p^*/2}$. Оценку для I_2 непосредственно заимствуем из [3]:

$$I_2 \le \int\limits_{|t| \ge \lambda\sqrt{p^*}} \frac{1}{|t|} \frac{1}{(1+t^2/p^*)^{p^*/2}}\, dt \le$$
$$\le \frac{1+\lambda^2}{\lambda^2} \int\limits_{u \ge \lambda^2} (1+u)^{-1-m/2}\, du =$$
$$= \frac{4\left(1+\lambda^2\right)}{\lambda^2 p} \left(1+\lambda^2\right)^{-p/4} .$$

Рассуждая аналогично [6] и используя интегрирование по частям, получаем

$$I_3 \le 2 \int\limits_{|t| \ge \lambda\sqrt{p^*}} \frac{1}{t} e^{-t^2/2}\, dt =$$
$$= \frac{2}{\lambda^2 p^*} \exp\left(-\frac{\lambda^2 p^*}{2}\right) - 2 \int\limits_{\lambda\sqrt{p^*}}^{\infty} t^{-3} e^{-t^2/2}\, dt \le$$
$$\le \frac{4}{\lambda^2 p} \exp\left(-\frac{\lambda^2 p}{4}\right) .$$

Объединяя оценки для I_1, I_2 и I_3, завершаем доказательство следующей леммы:

Лемма 2. *Для всех* $\lambda \in (0, \sqrt{3}-1)$ *и целых* $p > 1$

$$\sup_{x \in R} |F_p(x) - \Phi(x)| \le D_2(\lambda, p)\,,$$

где

$$D_2(\lambda, p) = \frac{2}{\pi} \left(\sqrt{\frac{\pi}{p}} \frac{1}{6} + \frac{2(1-\lambda)}{p(2-2\lambda-\lambda^2)^2} + \right.$$
$$\left. + \frac{(1+\lambda^2)}{\lambda^2 p} \left(1+\lambda^2\right)^{-p/4} + \frac{1}{\lambda^2 p} \exp\left(-\frac{\lambda^2 p}{4}\right) \right) .$$

Покажем, что оценка леммы 2 точнее оценки леммы 1, т. е.

$$D(\lambda, p) = \min\left(D_1(\lambda, p), D_2(\lambda, p)\right) = D_2(\lambda, p)\,. \quad (6)$$

Действительно, поскольку

$$\frac{3+\lambda}{3} > 1$$

и

$$\frac{3-\lambda}{3+\lambda} < 1$$

для положительных λ, равенство (6) верно.

В табл. 1 для некоторых значений p даны величины $\min\limits_{\lambda} D(\lambda, p)$ и соответствующие значения λ, при которых достигается минимум.

Таблица 1 Минимумы $D(\lambda, p)$ при фиксированном p

p	λ	$\min D$
10	0,551	0,460
30	0,517	0,104
50	0,483	0,0550
100	0,416	0,0276
500	0,244	0,0094

Литература

1. *Hall P., Marron J. S., Neeman A.* Geometric representation of high dimension, low sample size data // J. Royal Statistical Soc. Series B, 2005. Vol. 67. P. 427–444.
2. *Ulyanov V. V., Wakaki H., Fujikoshi Y.* Berry–Esseen bound for high dimensional asymptotic approximation of Wilks' lambda distribution // Statist. Probab. Lett., 2006. Vol. 76. No. 12. P. 1191–1200.
3. *Ulyanov V. V., Christoph G., Fujikoshi Y.* On approximations of transformed chi-squared distributions in statistical applications // Siber. Math. J., 2006. Vol. 47. No. 6. P. 1154–1166.
4. *Konishi S.* Asymptotic expansions for the distributions of functions of a correlation matrix // J. Multivariate Analysis, 1979. Vol. 9. No. 2. P. 259–266.
5. Справочник по специальным функциям / Под ред. М. Абрамовица и И. Стиган. — М.: Наука, 1979.
6. *Dobric V., Ghosh B. K.* Some analogs of the Berry–Esseen bounds for first-order Chebyshev–Edgeworth expansions // Statist. Decisions, 1996. Vol. 14. No. 4. P. 383–404.