

ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.233.22

К. Андреа, П.О. Смирнов, Г.Л. Шевляков

ДВУМЕРНЫЙ БОКСПЛОТ НА ОСНОВЕ ВЫСОКОЭФФЕКТИВНЫХ РОБАСТНЫХ ОЦЕНОК МАСШТАБА И КОРРЕЛЯЦИИ

На основе новых «быстрых» высокоэффективных и робастных FQn -оценок масштаба и корреляции предложен двумерный боксплот, ориентированный на визуализацию эллиптически распределенных данных. Обоснован выбор этих оценок и исследована вычислительная сложность алгоритма построения боксплота.

Ключевые слова: визуализация, двумерный боксплот, робастность.

В [1] предложен новый способ визуализации данных двумерными FQn -боксплотами, где показано, что FQn -боксплоты более эффективны для данных, распределенных по двумерному нормальному закону, чем другой хорошо зарекомендовавший себя непараметрический боксплот – бэгплот (bagplot) [2].

В разведочном анализе существует ряд методов первоначальной обработки одномерных данных. В настоящей статье рассматриваются проблемы анализа и представления двумерных данных. Отправной точкой в нашем обсуждении служит одномерный боксплот Тьюки [3], изображенный на рис. 1: он представляет собой прямоугольник с высотой, равной выборочному межквартильному расстоянию с отмеченной медианой в качестве оценки параметра положения и так называемыми «усами», зависящими от экстремальных порядковых статистик выборки.

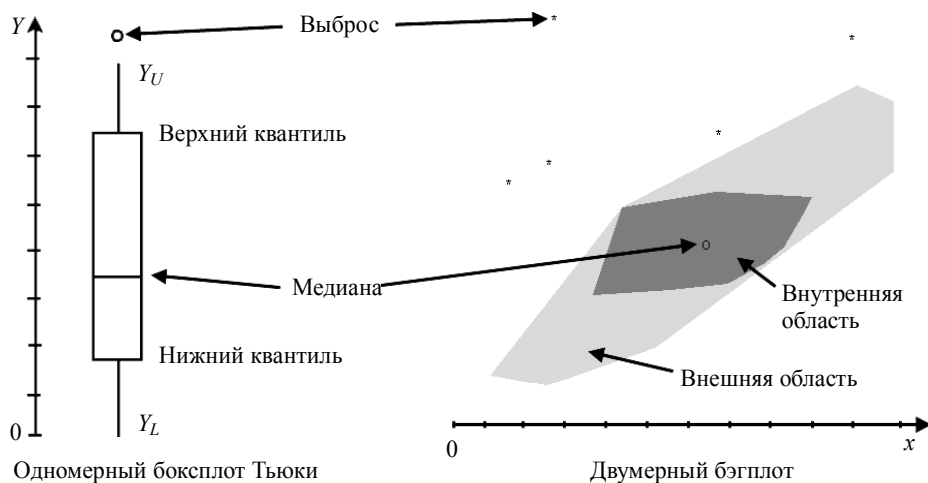


Рис. 1 Классический одномерный боксплот Тьюки и двумерный бэгплот

Двумерный боксплот является статистическим средством представления данных на плоскости. Он отображает информацию о параметрах положения, масштаба, асимметрии, «хвостах» и выбросах распределения (см. рис. 1). Для полноты описания двумерного распределения используется дополнительная характеристика статистической взаимосвязи между случайными величинами – коэффициент корреляции.

1. Робастные оценки параметра положения и масштаба

Использование робастных оценок параметров распределения (положения, масштаба, корреляции) в боксплотах является традиционным; при этом желательно, чтобы эти оценки были возможно более эффективными. В [4] проводится анализ десяти различных оценок параметров положения двумерных выборок: показано, что наилучшей оценкой является пространственная медиана (spatial median), минимизирующая сумму модулей невязок. Алгоритм вычисления пространственной медианы реализован в среде R; подробное его описание можно найти в [5].

Задача оценивания параметров масштаба занимает второе по значению место после оценивания параметров положения распределения. В общем случае, оценка параметра масштаба определяется статистикой

$$S_n : S_n(ax_1, \dots, ax_n) = |a| S_n(x_1, \dots, x_n).$$

Классической оценкой масштаба является стандартное отклонение, однако при наличии выбросов в выборке стандартное отклонение – неробастная и неэффективная оценка.

Методы робастной статистики предлагают более устойчивые статистические оценки для случаев, когда в выборке данных присутствуют выбросы, в частности робастная, высокоэффективная, но вычислительно сложная Qn -оценка масштаба [6]. В [7] предложена «быстрая» робастная высокоэффективная FQn -оценка масштаба, основанная на аппроксимации функции влияния Qn -оценки. Показано, что максимальная эффективность предложенной FQn -оценки достигает 96%, а минимальное возможное ее значение не опускается ниже уровня 81% на нормальном распределении, при этом их пороговая точка (breakdown point) достигает 50%. Вычисление модифицированной FQn -оценки по выборке $\{x_1, \dots, x_n\}$ производится по формуле

$$FQ_n(x) = 1,483 \text{ MAD} \left(1 - \frac{Z_0 - \frac{n}{\sqrt{2}}}{Z_2} \right), \quad (1)$$

где
$$Z_k = \sum u_i^k e^{-\frac{u_i^2}{2}}, \quad u_i = \frac{x_i - \text{med } x}{1,483 \cdot \text{MAD}},$$

$\text{med } x$ – выборочная медиана, а $\text{MAD} = 1,483 \text{ med}|x_i - \text{med } x|$ – медианное абсолютное отклонение. Поправочный коэффициент 1,483 обеспечивает состоятельность оценки MAD и FQn на нормальных распределениях. Параметр масштаба, определяемый в уравнении (1), является одношаговой M -оценкой [7]. В среде R создан пакет {fastqn}, где реализован алгоритм вычисления быстрых FQn -оценок масштаба.

2. Робастное оценивание коэффициента корреляции

Классическая оценка корреляции двумерных данных дается выборочным коэффициентом корреляции Пирсона

$$r = \sum (x_i - \bar{x})(y_i - \bar{y}) / \left(\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2 \right)^{1/2},$$

где $\bar{x} = \frac{\sum x_i}{n}$ и $\bar{y} = \frac{\sum y_i}{n}$. Как отмечено выше, классические оценки положения и

масштаба, такие, как арифметическое среднее и дисперсия, неработоспособны в присутствии загрязненных данных. Различные робастные оценки коэффициента корреляции рассматриваются в [8, 9].

В [1] коэффициент корреляции вычисляется по методу минимального определителя ковариационной матрицы [10]. Ковариационная матрица, полученная по методу минимального определителя, является аффинно-эквивариантной оценкой, обладающей высокой эффективностью 88% на двумерном нормальном распределении и пороговой точкой около 25%. Применение метода минимального определителя ковариационной матрицы для построения FQn -боксплота обусловлено его устойчивым поведением в присутствии выбросов, а также его аффинно-эквивариантным свойством. Наряду с робастным коэффициентом по методу минимального определителя в нашей работе используется FQn робастная оценка коэффициента корреляции [11]:

$$r_{FQn} = \frac{FQ_n^2(x+y) - FQ_n^2(x-y)}{FQ_n^2(x+y) + FQ_n^2(x-y)}, \quad (2)$$

где векторы x и y центрированы соответствующими компонентами пространственной медианы и нормированы соответствующими FQn -оценками.

В работе [8] проведено сравнение различных методов по вычислению коэффициента корреляции. В результате такого исследования наилучшим среди робастных оценок коэффициента корреляции является r_{FQn} .

3. Алгоритм построения двумерного FQn -боксплота

Приведем алгоритм построения FQn -боксплота [1]:

1) Заданные точки (x, y) на плоскости преобразуются в новую систему координат (x', y') , определяемую главными осями эллипса равной вероятности для нормального распределения с центром (x_c, y_c) : $(x_c, y_c) = \text{spatial median}(x, y)$ [7],

$$\begin{aligned} x' &= (x - x_c) \cos \alpha + (y - y_c) \sin \alpha \\ y' &= -(x - x_c) \sin \alpha + (y - y_c) \cos \alpha \end{aligned} \quad (3)$$

где $\tan 2\alpha = (2r_{MCD} \hat{\sigma}_1 \hat{\sigma}_2) / (\hat{\sigma}_1^2 - \hat{\sigma}_2^2)$.

Главные оси (x', y') эллипса рассеивания двумерного нормального распределения пропорциональны оценкам масштаба, а угол поворота α главных осей зависит также и от оценки коэффициента корреляции.

2) Выбирается половина ближайших по расстоянию Махаланобиса точек к центру. Выпуклая оболочка этих точек определяет внутреннюю область боксплота. Расстояние Махаланобиса точек от центра μ определяется

$$D_M = \sqrt{(x - \mu)^T S^{-1} (x - \mu)},$$

где S – ковариационная матрица, имеющая следующий вид:

$$S = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 \\ \hat{\rho}\hat{\sigma}_1\hat{\sigma}_2 & \hat{\sigma}_2^2 \end{pmatrix}, \quad (4)$$

где $\hat{\sigma}_1 = FQn(x')$, $\hat{\sigma}_2 = FQn(y')$ и $\hat{\rho} = r_{FQn}$.

3) Внешняя область представляет собой выпуклую оболочку множества точек, которые не попали во внутреннюю область и при этом находятся в пределах эллипса, вписанного в прямоугольник со сторонами, параллельными осям x' и y' с левой и правой границами:

$$\begin{aligned} x'_L &= \max\{x_{(1)}, x_c - aFQn(x')\}, & y'_L &= \max\{y_{(1)}, y_c - aFQn(y')\}, \\ x'_R &= \min\{x_{(n)}, x_c + aFQn(x')\}, & y'_R &= \min\{y_{(n)}, y_c + aFQn(y')\}, \end{aligned} \quad (5)$$

где a – коэффициент, значение которого выбирается исходя из модели распределения данных.

4) Точки, лежащие за границей внешней области, рассматриваются как выбросы (рис. 2).

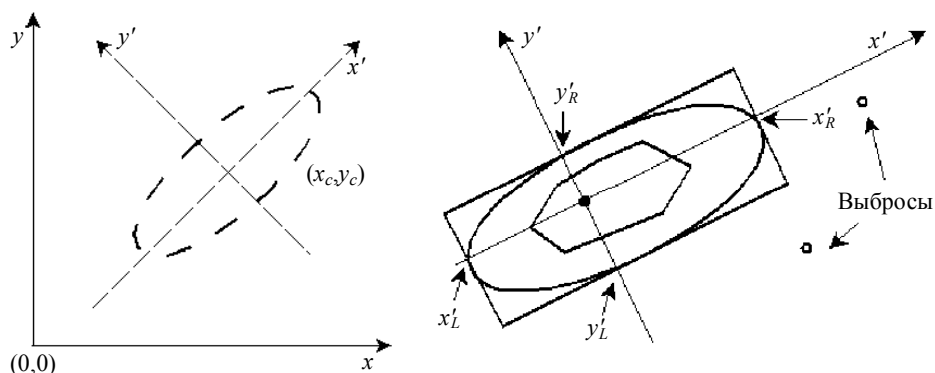


Рис. 2 Построение FQn -боксплота

4. Вычислительная сложность алгоритма построения FQn -боксплота

Алгоритмическая сложность двумерного FQn боксплота составляет

$$O(n\sqrt{\log(n)}),$$

где n – размер выборки. Ниже приводится список операций над выборкой и их соответственная алгоритмическая сложность:

1) Вычисление ковариационной матрицы минимального определителя является линейной функцией от числа входных данные, т.е. $O(n)$.

2) Сложность перевода точек из одной системы координат в другую составляет $O(n)$.

3) Пространственная медиана вычисляется итеративным алгоритмом $O(cn)$, где в худшем случае $c = 500$ (c – число итераций).

4) Вычисление приближенной FQn -оценки масштаба составляет $O(n)$.

5) Алгоритм вычисления выпуклой оболочки точек составляет $O(n\sqrt{\log(n)})$, когда точки на плоскости распределены по нормальному закону.

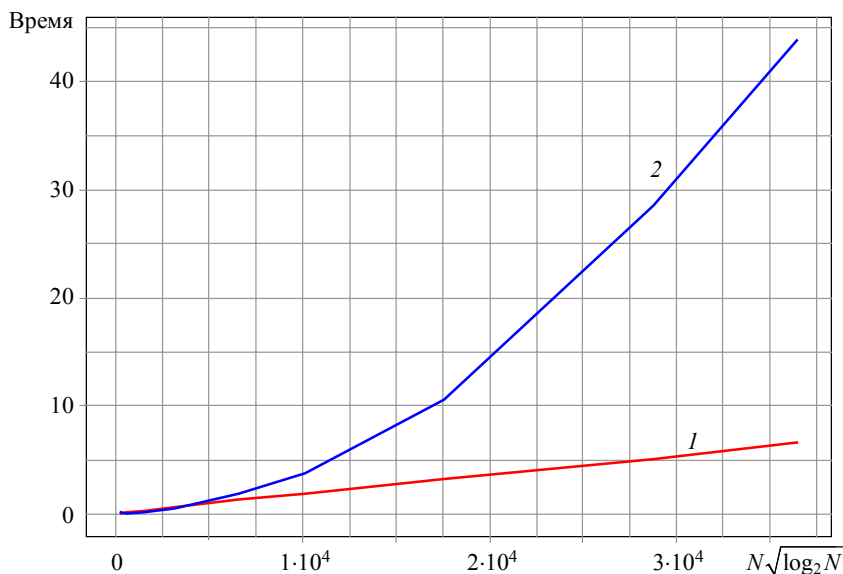


Рис. 3. Сравнение времени вычисления FQn -боксплота (1) и бэглота (2) при увеличении размера выборки (преобразование над N : $N\sqrt{\log N}$)

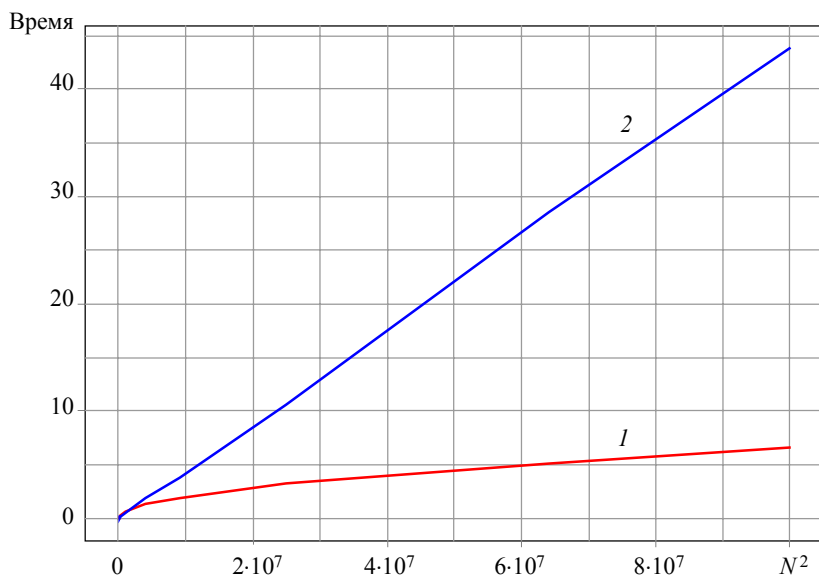


Рис. 4. Сравнение времени вычисления FQn -боксплота (1) и бэглота (2) при увеличении размера выборки (преобразование над N : N^2)

Итоговая сложность построения нашего двумерного FQn -боксплота составляет $O(n\sqrt{\log(n)} + 503n)$. Исследование сравнения времени вычисления алгоритмов FQn -боксплота и бэглота показало, что FQn -боксплот работает намного быстрее. Исследование проводилось по методу Монте-Карло с увеличением размера дву-

мерной выборки, распределенной по нормальному закону, от 100–10000. Измерение времени вычисления для каждого размера выборки оценивалось арифметическим средним по времени из 10 разных выборок заданного размера. Вычисления проводились на ЭВМ Intel(R) Core(TM) i7-2620M, 2.7 GHz, Windows 7 64bit.

На рис. 3 легко увидеть линейную зависимость времени вычисления FQn -боксплота от преобразованного размера выборки $N\sqrt{\log N}$. Применение линейной регрессии свидетельствует о том, что такая зависимость приближается к прямой (коэффициенты наклона и пересечения 0,001 и $-3,5$) с очень высокой степенью достоверности ($R^2 = 0,99$). Бэблот в свою очередь имеет вычислительную сложность $O(N^2)$. Как можно увидеть из рис. 4, применение линейной регрессии для бэбплота с высокой степенью достоверности ($R^2 = 0,99$) дает квадратичную зависимость времени вычисления от преобразованного размера выборки (коэффициенты наклона и пересечения $4,4 \cdot 10^{-7}$ и $3,4 \cdot 10^{-2}$).

Заключение

Предложенный FQn -боксплот построен на высокоэффективных оценках масштаба и корреляции, что обеспечивает его превосходство над бэблотом. Рассматриваемые в литературе двумерные боксплоты делятся на два больших класса: ориентированные на данные и модельно-ориентированные боксплоты. Предложенный нами FQn -боксплот относится к классу модельно-ориентированных боксплотов, а именно ориентирован на модель двумерного нормального распределения. Превосходство FQn -боксплота над бэблотом при отбраковке выбросов в случае двумерных нормально распределенных данных объясняется применением робастных высокоэффективных FQn -оценок параметров масштаба и коэффициента корреляции.

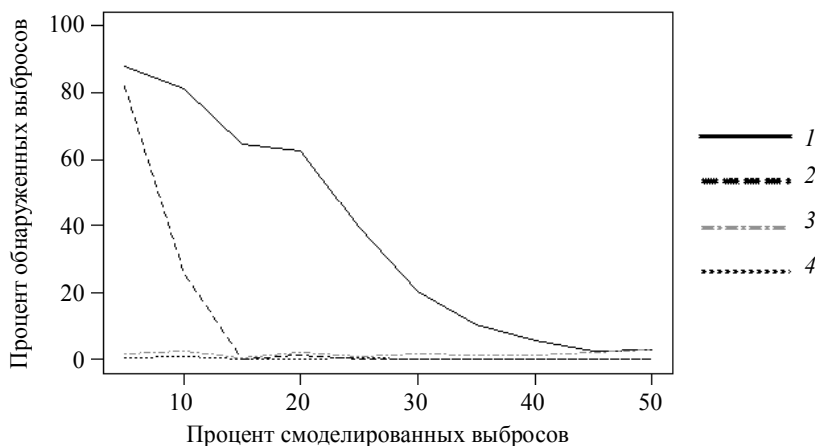


Рис. 5 Результаты отбраковки выбросов: правильно обнаруженные выбросы — FQn -боксплот (1) и бэблот (2); неправильно — FQn -боксплот (3) и бэблот (4)

На рис. 5 проиллюстрировано поведение FQn -боксплота при увеличении количества выбросов — оно является более устойчивым по сравнению с бэблотом. Рис. 5 отражает значения чувствительности и специфичности, полученные для выборки, удовлетворяющей модели Тьюки — Хьюбера в виде засорения типа «сдвиг»

[12] с различными значениями параметра засорения ε (вероятность появления выбросов в данных или процент смоделированных выбросов). Более подробное описание исследования мощности отбраковки выбросов при помощи двумерного FQn -боксплота можно найти в [1].

ЛИТЕРАТУРА

1. Андреа К., Лаврентьева Г.М., Смирнов П.О., Шевляков Г.Л., Визуализация данных двумерными fqn -боксплотами // Высокие технологии, фундаментальные исследования, экономика. Т.1: Сб. статей XII Междунар. научно-практ. конф. «Фундаментальные и прикладные исследования, разработка и применение высоких технологий в промышленности». СПб.: Изд-во Политехн. ун-та, 2011. С. 59–65.
2. Rousseeuw P.J., Ruts I, Tukey J.W. The bagplot: A bivariate boxplot // The American Statistician. 1999. V. 53. P. 382–387.
3. Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ. М.: Мир, 1981.
4. Masse J.C., Plante J.F. A Monte Carlo study of the accuracy and robustness of ten bivariate location estimators // Computational Statistics & Data Analysis. 2003. V. 42. P. 1–26.
5. Vardi Y. and Zhang C.H. The multivariate L1-median and associated data depth // PNAS. 1999. V. 97. P. 1423–1426.
6. Croux C., Rousseeuw P.J. Time-efficient algorithms for two highly robust estimators of scale // Computational Statistics. 1992. V. 1. P. 411–428.
7. Смирнов П.О., Шевляков Г.Л. Приближение оценки Qn параметра масштаба с помощью быстрых M -оценок // Вестник Сибирского государственного аэрокосмического университета. 2010. Т. 5 (31). С. 83–85.
8. Shevlyakov G., Smirnov P. Robust estimation of the correlation coefficient: An attempt of survey // Austrian J. Statistics. 2011. V. 40. P. 147–156.
9. Shevlyakov G.L., Vilchevski N.O. Robustness in data analysis: criteria and methods. Utrecht: VSP, 2002. 315 p.
10. Hubert M., Rousseeuw P.J., Van Aelst S. High-breakdown robust multivariate methods // Statistical Science. 2008. V. 23. P. 92–119.
11. Bernholt T., Fischer P. The complexity of computing the MCD-estimator // Theoretical Computer Science. 2004. V. 326. P. 383–398.
12. Filzmoser P., Identification of Multivariate Outliers: A performance study // Austrian J. Statistics. 2005. V. 34. P. 127–138.

Андреа Клитон

Смирнов Павел Олегович

Шевляков Георгий Леонидович

Санкт-Петербургский государственный политехнический университет

E-mail: kliton.andrea@gmail.com; s.paul@mail.ru; gshevlyakov@yahoo.com

Поступила в редакцию 4 мая 2012 г.

Andrea Kliton, Smirnov Pavel O., Shevlyakov Georgy L. (St. Petersburg State Polytechnical University). **A bivariate boxplot based on robust highly efficient estimators of scale and correlation.**

Keywords: visualization, bivariate boxplot, robustness.

A bivariate model-based boxplot based on fast highly efficient and robust FQn -estimates of scale and correlation is proposed. The choice of parameters is motivated by their high performance and is based on the state-of-the-art methods. It is shown that FQn -boxplot has a better speed performance over the conventional boxplot.