ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.6

А.В. Бобров, Е.А. Перепелкин

ПОСТРОЕНИЕ РОБАСТНЫХ ОЦЕНОК СРЕДНИХ ЗНАЧЕНИЙ И ВАРИАЦИЙ ДВУМЕРНЫХ ДАННЫХ НА ОСНОВЕ СПЕКТРАЛЬНОЙ МАТРИЧНОЙ НОРМЫ

Описываются алгоритмы построения робастных оценок средних значений и вариаций двумерных данных, полученные на основе спектральной матричной нормы. Рассматриваются алгоритмы для дискретных и непрерывных данных. Анализируются результаты тестовых расчетов.

Ключевые слова: *анализ данных*, *среднее значение*, *вариация*, *робастность*.

Теория робастных оценок составляет одно из наиболее важных и актуальных направлений в методах обработки данных [1–3]. Необходимость построения робастных оценок связана с целым рядом причин: ограниченным объемом выборки, пропуском данных, ошибками в записи данных и др. Один из подходов к построению робастных оценок основан на решении задач оптимизации. Пусть данные измерений представлены числовыми значениями x_1, x_2, …, x_n. Классическая оценка среднего значения

$$c = \sum_{i=1}^{n} x_i w_i ,$$

где w_i – весовые коэффициенты,

$$\sum_{i=1}^{n} w_i = 1 , \quad w_i \ge 0 .$$

Эта оценка является решением задачи оптимизации

$$c = \arg\min_s \sum_{i=1}^{n} (s - x_i)^2 w_i .$$

Робастную оценку можно получить, решая задачу оптимизации

$$c = \arg\min_s \sum_{i=1}^{n} | s - x_i |^p \ w_i .$$

где $1 \le p < 2$. Такие оценки принадлежат к классу M-оценок и называются L_p-оценками [3].

В данной работе описываются алгоритмы построения робастных оценок средних значений и вариаций двумерных данных, полученные на основе спектральной матричной нормы. Рассматриваются алгоритмы для дискретных и непрерывных данных. Приводятся и анализируются результаты тестовых расчетов.

1. Дискретные данные

Пусть числовые данные представлены в виде матрицы

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}.$$

Обычно вычисляют взвешенное среднее значение c и вариацию данных d в матрице A:

$$c = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij} w_{ij} , \qquad d = \sum_{i=1}^{n}\sum_{j=1}^{m} (c - a_{ij})^2 w_{ij} ,$$

где w_{ij} – весовые коэффициенты,

$$\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij} = 1 , \qquad w_{ij} \geq 0 .$$

Заметим, что значение c является решением задачи оптимизации

$$c = \arg\min_{s} \left\| B(s) \right\|_F ,$$

где

$$\left\| B(s) \right\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{m} (b_{ij}(s))^2 \right)^{1/2}$$

– норма Фробениуса матрицы $B(s)$ с элементами $b_{ij}(s) = (s - a_{ij})\sqrt{w_{ij}}$. При этом

$$d = \left\| B(c) \right\|_F^2 .$$

В вычислительной математике, методах обработки данных [4, 5] наряду с нормой Фробениуса применяются и другие матричные нормы, например гельдеровы нормы:

$$\left\| B \right\|_1 = \max_j \sum_{i=1}^{n} \left| b_{ij} \right| , \qquad \left\| B \right\|_2 = \sigma_{\max}(B) , \qquad \left\| B \right\|_\infty = \max_i \sum_{j=1}^{m} \left| b_{ij} \right| ,$$

где $\sigma_{\max}(B)$ – максимальное сингулярное число матрицы B . В расчетах на реальных данных часто применяется 2-норма (спектральная норма), поскольку данная норма обладает свойством грубости (робастности) по отношению к изменениям в элементах матрицы [5]. Изменения (возмущения) в матрице данных возможны по ряду причин: технические ошибки при записи данных, отсутствие данных и др. Рассмотрим влияние возмущений в матрице данных на спектральную норму матрицы $B(s)$.

Сингулярные числа матрицы A размером $n \times m$, упорядоченные по невозрастанию, обозначим $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_{\min(n,m)}(A)$. Справедливы следующие утверждения [5].

Утверждение 1. Пусть матрицы P , Q , R размером $n \times m$ связаны соотношением $P = Q + R$. Тогда

$$\left| \sigma_i(P) - \sigma_i(Q) \right| \leq \sigma_1(R) , \quad i = 1, \ldots, \min(n, m) .$$

Утверждение 2. Пусть матрица Q получена из матрицы P вычеркиванием строки или столбца. Тогда $\sigma_1(P) \geq \sigma_1(Q) \geq \sigma_2(P)$.

Рассмотрим два вида возмущений в матрице данных: ошибки в записи элементов матрицы и пропуски строк или столбцов.

Утверждение 3. Пусть две матрицы данных A и $\overline{A}$ отличаются k элементами, т.е.

$$\overline{A} = A + \sum_{i=1}^{k} A_i \delta_i ,$$

где в матрицах A_i все элементы равны нулю, за исключением одного, равного единице. Составим вектор возмущений $\delta = [\delta_1, \ ..., \ \delta_k]$. Тогда для любого s справедлива оценка

$$\left| \left\| \overline{B}(s) \right\|_2 - \left\| B(s) \right\|_2 \right| \leq \left\| \delta \right\|_2 ,$$

где $\left\| \delta \right\|_2 = \left(\sum_{i=1}^{k} \delta_i^{\,2} \right)^{1/2}$.

Доказательство. Обозначим через Δa_{ij} элементы матрицы

$$\Delta A = \sum_{i=1}^{k} A_i \delta_i .$$

Тогда $\overline{b}_{ij}(s) = b_{ij}(s) - \Delta a_{ij} \sqrt{w_{ij}}$. В соответствии с утверждением 1

$$\left| \sigma_1\left(\overline{B}(s)\right) - \sigma_1\left(B(s)\right) \right| \leq \sigma_1(R) ,$$

где матрица R состоит из элементов $\Delta a_{ij} \sqrt{w_{ij}}$. Справедливо неравенство [4]

$$\sigma_1(R) \leq \left\| R \right\|_F .$$

Из структуры матрицы R следует неравенство

$$\left\| R \right\|_F \leq \left\| \delta \right\|_2 .$$

Таким образом, для любого s справедлива оценка

$$\left| \sigma_1\left(\overline{B}(s)\right) - \sigma_1\left(B(s)\right) \right| \leq \left\| \delta \right\|_2 .$$

Утверждение доказано.

Утверждение 4. Пусть в результате удаления строки или столбца в матрице данных A получена матрица $\overline{A}$. Тогда для любого s справедлива оценка

$$\left\| B(s) \right\|_2 \geq \left\| \overline{B}(s) \right\|_2 \geq \sigma_2(B(s)) .$$

Доказательство. Удаление строки или столбца в матрице A означает удаление соответствующей строки или столбца в матрице $B(s)$. На основании утверждения 2 мы можем записать

$$\sigma_1(B(s)) \geq \sigma_1(\overline{B}(s)) \geq \sigma_2(B(s)) .$$

Утверждение доказано.

Мы предлагаем для оценки среднего значения и вариации данных в двумерном массиве использовать спектральную норму. Среднее значение определим как решение задачи оптимизации

$$c = \arg\min_s \|B(s)\|_2 .$$

Вариацию данных будем вычислять по формуле

$$d = \|B(c)\|_2^2 .$$

Определить значения c и d достаточно просто, применяя системы компьютерной математики, например MATLAB или SCILAB.

Пример 1. Тестовые расчеты проводились в системе MATLAB на данных модульно-рейтинговой системы квалиметрии учебной деятельности студентов АлтГТУ. Матрица данных является ведомостью студенческой группы с семестровыми рейтингами студентов в 100-балльной шкале по предметам семестра. Заданы веса предметов. Необходимо определить рейтинг группы в целом и вариацию рейтинга в группе. Оценки рейтинга группы, полученные на основе нормы Фробениуса и спектральной нормы матрицы $B(s)$, соответственно равны 65 и 64, оценки вариации рейтинга в группе равны $19{,}2$ и $14{,}7$.

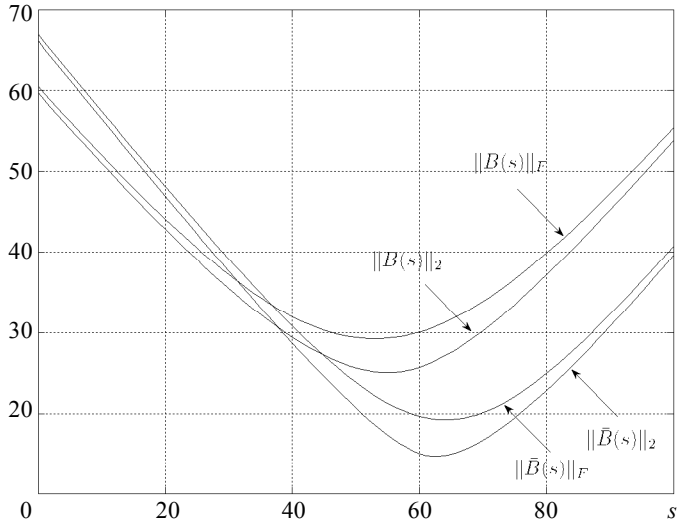

Рис. 1. Зависимость норм матриц $B(s)$ и $\bar{B}(s)$ от s в примере 1

Предположим, что в ведомости не проставлены 10 % оценок. Соответствующие элементы матрицы $\bar{A}$ полагаются равными нулю. Оценки рейтинга группы равны 54 и 56, оценки вариации рейтинга в группе равны $29{,}3$ и $25{,}1$ соответственно для нормы Фробениуса и спектральной нормы матрицы $\bar{B}(s)$.

Таким образом, оценка рейтинга группы изменилась на 16,9 % и 12,5 % при расчете соответственно на основе нормы Фробениуса и спектральной нормы.

2. Непрерывные данные

Пусть имеется массив данных в виде значений функции двух аргументов

$$A = \{a(x,y) / (x,y) \in D = [x_{\min}, x_{\max}] \times [y_{\min}, y_{\max}]\} .$$

Например, это может быть фрагмент цифровой фотографии, данные геофизических измерений и т.д. Необходимо построить оценки среднего значения и вариации данных в области D, грубые по отношению к ошибкам в данных.

Среднее значение и вариация определяются в виде интегралов

$$c = \int_{x_{\min}}^{x_{\max}} \int_{y_{\min}}^{y_{\max}} a(x,y) w(x,y) dxdy , \quad d = \int_{x_{\min}}^{x_{\max}} \int_{y_{\min}}^{y_{\max}} (c - a(x,y))^2 \, w(x,y) dxdy ,$$

где $w(x,y)$ – весовая функция,

$$\int_{x_{\min}}^{x_{\max}} \int_{y_{\min}}^{y_{\max}} w(x,y) dxdy = 1 , \qquad w(x,y) \ge 0 , \qquad (x,y) \in D .$$

Введем функцию трех аргументов $b(s,x,y) = (s - a(x,y))\sqrt{w(x,y)}$. Тогда

$$d = \|b(c,x,y\|^2 ,$$

где

$$\|b(s,x,y)\| = \left(\int_{x_{\min}}^{x_{\max}} \int_{y_{\min}}^{y_{\max}} b(s,x,y)^2 dxdy \right)^{1/2}$$

– норма функции $b(s,x,y)$ как функции двух аргументов x и y в пространстве $L_2(D)$. В свою очередь, значение c есть решение задачи оптимизации

$$c = \arg\min_s \|b(s,x,y)\| .$$

Норму функции $b(s,x,y)$ в пространстве $L_2(D)$ можно приближенно вычислить, выполнив дискретизацию множества D конечным числом точек (x_i, y_j), $i = 1,...,n$, $j = 1,...,m$. Например, при равномерной дискретизации по каждой из переменных

$$x_i = x_{\min} + \frac{2i-1}{2} \Delta_x , \qquad y_j = y_{\min} + \frac{2j-1}{2} \Delta_y ,$$

где

$$\Delta_x = \frac{x_{\max} - x_{\min}}{n} , \qquad \Delta_y = \frac{y_{\max} - y_{\min}}{m} ,$$

оценка нормы $b(s,x,y)$ имеет следующий вид:

$$\|b(s,x,y)\| \approx \sqrt{\Delta_x \Delta_y} \, \|B(s)\|_F .$$

Здесь

$$\|B(s)\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} (b_{ij}(s))^2 \right)^{1/2}$$

– норма Фробениуса матрицы $B(s)$ с элементами $b_{ij}(s) = s - a(x_i, y_j)$.

Таким образом, среднее значение данных в области D может быть найдено как решение задачи оптимизации

$$c = \arg\min_s \|B(s)\|_F .$$

При этом вариация данных будет равна

$$d = \Delta_x \Delta_y \|B(c)\|_F^2 .$$

От нормы Фробениуса перейдем к спектральной норме матрицы $B(s)$. Среднее значение определим как решение задачи оптимизации

$$c = \arg\min_s \|B(s)\|_2 .$$

Вариацию данных будем вычислять по формуле

$$d = \Delta_x \Delta_y \|B(c)\|_2^2 .$$

Пример 2. Пусть анализируемые данные представлены функцией

$$a(x, y) = 1,2(x-5)^2 + 1,5(y+7)^2 + 2,7$$

в области $D = [3,7] \times [-10,-5]$. На рис. 2 показаны значения $a(x, y)$ в области D. Расчеты проводились при $n = 100$, $m = 100$. На рис. 3 показана зависимость $\|B(s)\|_F$ и $\|B(s)\|_2$ от s для исходных данных. Значения

$$c = \arg\min_s \|B(s)\|_F , \quad c = \arg\min_s \|B(s)\|_2$$

совпадают и равны 7,8. Значения

$$d = \Delta_x \Delta_y \|B(c)\|_F^2 , \quad d = \Delta_x \Delta_y \|B(c)\|_2^2$$

равны соответственно 12,5 и 13,81.

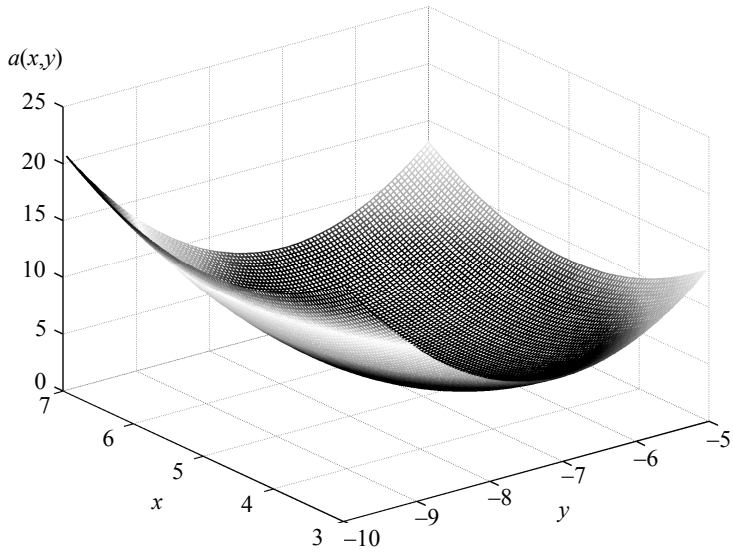

Рис. 2. Исходные данные

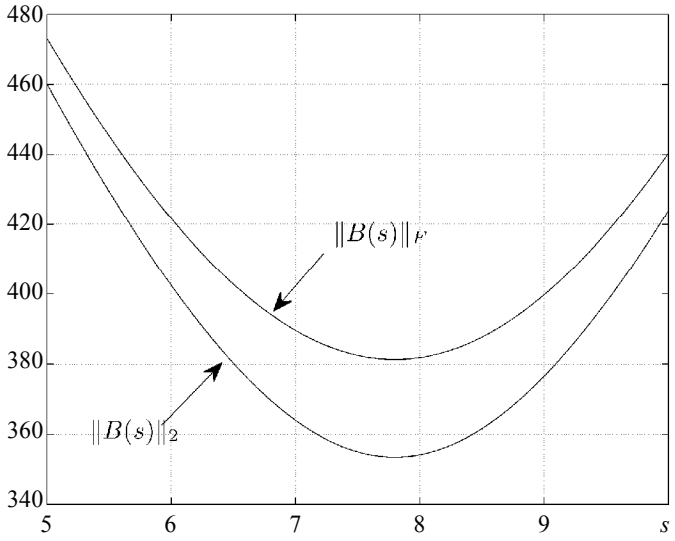

Рис. 3. Зависимость норм матрицы $B(s)$ от s для исходных данных

Пусть в анализируемых данных пропущены 20 % значений. Соответствующие значения данных показаны на рис. 4. Зависимость $\|B(s)\|_F$ и $\|B(s)\|_2$ от s при пропущенных данных показана на рис. 5. В этом случае оценки средних значений и вариаций данных существенно отличаются при расчете на основе нормы Фробениуса и спектральной нормы матрицы $B(s)$. Значение c, полученное на основе нормы Фробениуса, равно 6,39, на основе спектральной нормы – 7,48. Значения

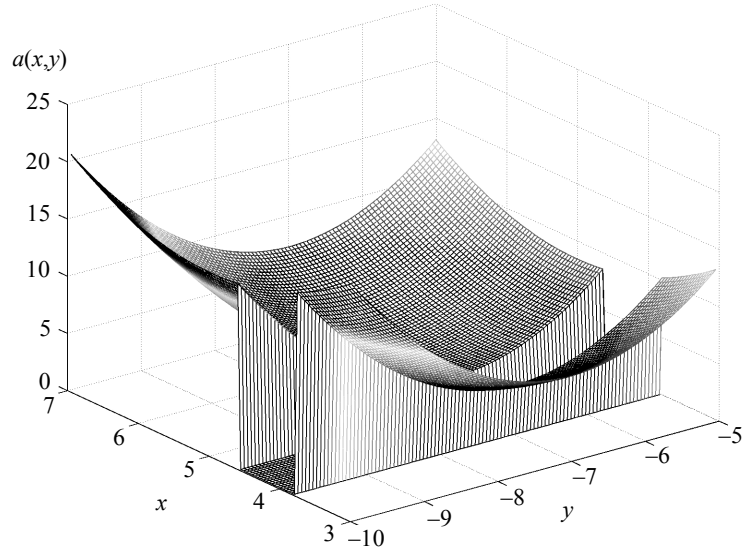

Рис. 4. Данные с пропущенными значениями

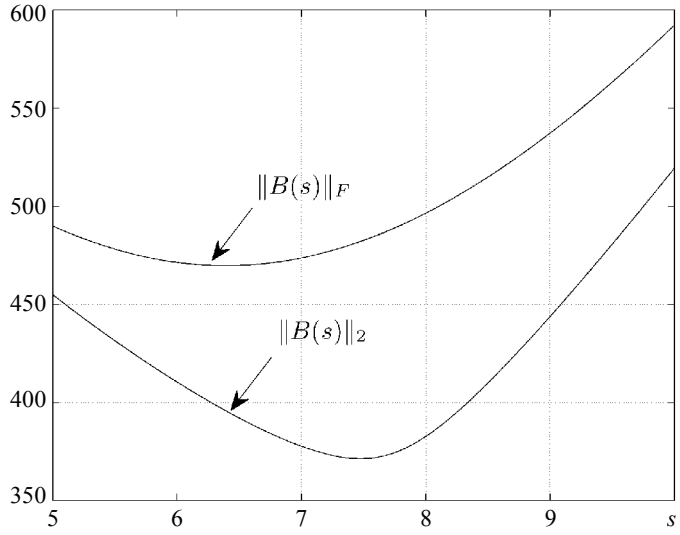

Рис. 5. Зависимость норм матрицы $B(s)$ от s для данных с пропущенными значениями

d равны соответственно 22,07 и 13,81. Таким образом, оценка среднего значения, полученная на основе нормы Фробениуса, изменилась на 10,5 %, на основе спектральной нормы – на 4,1 %. Оценки вариаций изменились соответственно на 51,8 и 10,5 %.

Заключение

В статье описан метод расчета средних значений и вариаций двумерных дискретных и непрерывных данных на основе спектральной матричной нормы. Получены неравенства, позволяющие оценить влияние ошибок в матрице данных на результаты расчетов средних значений и вариаций данных. Результаты тестовых расчетов подтверждают грубость оценок средних значений и вариаций, полученных на основе спектральной нормы, по отношению к ошибкам в данных.

ЛИТЕРАТУРА

1. *Смоляк С.А.*, *Титаренко Б.П.* Устойчивые методы оценивания. М.: Статистика, 1980. 208 с.
2. *Хьюбер Дж.П.* Робастность в статистике. М.: Мир, 1984. 304 с.
3. *Крянев А.В.*, *Лукин Г.В.* Математические методы обработки неопределенных данных. М.: Физматлит, 2003. 216 с.
4. *Воеводин В.В. Кузнецов Ю.А.* Матрицы и вычисления. М.: Наука, 1984. 320 с.
5. *Лоусон Ч.*, *Хенсон Р.* Численное решение задач метода наименьших квадратов. М.: Наука, 1986. 232 с.

Бобров Александр Валерьевич
Перепелкин Евгений Александрович
Алтайский государственный технический университет
им. И.И. Ползунова
E-mail; 22bav@mail.ru eap@list.ru

Поступила в редакцию 30 октября 2011 г.

Bobrov Alexandr V., *Perepelkin Evgeniy A.* (Polzunov Altai State Technical University). **Construction of robust estimates of meano values and variations of two-dimensional data on the basis of the spectral matrix norm**.

Keywrds: data analysis, mean value, variation, robustness.

The problem of estimating of mean values and variations of two-dimensional data is considered. Data can be represented as a matrix in discrete case, or as a function of two variables in the continuous case. Estimate of the mean value of discrete data is found as solution of the optimization problem with objective function in the form of the spectral norm of a matrix. In the continuous case the problem is reduced to the discrete case by sampling the domain of function. The inequalities to assess the influencet of errors in the data matrix on the results of calculations of mean values and variations in data are derived. The results of test calculations confirm the robustness of the estimates of the mean values and variations derived from spectral norms with respect to data errors.