УДК 519.23

Об эквивалентности алгебраического и геометрического подходов к неискаженной визуализации многомерных данных

Калинкин А.А.
Алтайский государственный университет, г. Барнаул
kalinkin.7621@gmail.com

Аннотация

В работе рассматривается задача неискаженного изображения конфигурации объектов на плоскости по матрице попарных расстояний. Проводится сравнение классического метода многомерного шкалирования и предложенного ранее автором неитерационного геометрического алгоритма. Доказывается, что в условиях возможности построения неискаженного изображения оба метода дают результаты, эквивалентные с точностью до симметрий и поворотов. Приведен численный пример.

Ключевые слова: Неискаженное изображение, многомерные данные, визуализация статистических данных

1. Многомерное шкалирование

Задача визуализации многомерных данных по матрице их попарных различий в основном сводится к поиску конфигурации точек в пространстве низкой размерности (обычно двумерном), попарные расстояния между которыми максимально близки к исходным различиям. Классическим методом решения этой задачи является многомерное шкалирование. Пусть $D = \{d_{i,j}\}$ матрица попарных расстояний между n объектами. Поиск координат наилучшего их изображения осуществляется через матрицу двойного центрирования B, элементы которой вычисляются как:

$$b_{i,j} = -\frac{1}{2}\left(d_{i,j}^2 - \frac{1}{n}\sum_{j=1}^{n} d_{i,j}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{i,j}^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{i,j}^2\right), \quad i,j = 1,\ldots,n,$$

Известно, что матрица B связана с матрицей координат X соотношением

$$B = XX^T.$$

Основой метода служит теорема, которую в современной литературе часто связывают с именем У. Торгерсона. Однако справедливо будет отметить, что основные ее утверждения были впервые сформулированы и доказаны Г. Янгом и А. Хаусхолдером [1]. Торгерсон же переформулировал эти результаты в удобной для использования форме и разработал на их основе практический метод шкалирования [2].

Из главного теоретического результата [2] вытекает, что конфигурация точек $X_1,\ldots,X_n$ в евклидовом пространстве $\mathbb{R}^q (q \le n-1)$ такая, что

$$d_{i,j} = \sqrt{\sum_{s=1}^{q}(x_{i,s} - x_{j,s})^2}, \quad i,j = 1,\ldots,n,$$

существует тогда и только тогда, когда матрица B является положительно полуопределенной (то есть все собственные значения матрицы неотрицательны) и её ранг равен q.

В этом случае координаты точек X могут быть найдены как

$$X = F_q \cdot \sqrt{\Lambda_q},$$

где F_q – матрица из n строк и q столбцов, представляющих собой координаты собственных векторов, а $\Lambda_q = diag\{\lambda_1, \ldots, \lambda_q\}$ – диагональная матрица собственных значений матрицы B. Вычисления по этой формуле назовем матричным подходом к решению задачи.

Следствие 1. *Если ранг матрицы B равен 2, то данные объекты могут быть изображены на плоскости без искажений.*

Доказательство. По свойствам рангов матрицы, $rank(XX^T) \leq \min\{rank(X), rank(X^T)\}$, а $rank(X) = rank(X^T)$ (см. [3]). Следовательно, $rank(B) = rank(XX^T) = rank(X) = 2$. А значит, конфигурация точек может быть представлена в плоскости без каких-либо искажений расстояний, что и завершает доказательство. □

2. Геометрический подход

В работе автора [4] был предложен альтернативный, чисто геометрический алгоритм построения неискаженной конфигурации точек. В отличие от матричного подхода, он базируется на последовательном построении точек на плоскости.

Алгоритм работает следующим образом:

1. Первая точка помещается в начало координат, вторая — на оси абсцисс на расстоянии $d_{1,2}$.

2. Третья точка строится как произвольная точка пересечения окружностей радиусов $d_{1,3}$ и $d_{2,3}$ с центрами в уже построенных точках.

3. Аналогично по первых трем точкам можно построить все остальные точки по одной. Для этого строятся окружности с центрами в первой и третьей точках и радиусами $d_{1,k}$ и $d_{3,k}$ соответственно. В пересечении будет две точки. Из возможных альтернатив нужной будет та, у которой полученное расстояние до второй точки совпало с исходным. Существование правильной альтернативы следует из предположения о возможности полного построения неискаженной конфигурации.

Необходимым и достаточным условием возможности такого построения без искажений расстояний является равенство нулю объема тетраэдра с вершинами в любых четырех точках множества. Этот объем может быть рассчитан по длинам его ребер.

3. Эквивалентность методов

Поскольку решающим обстоятельством является не конкретное положение точек, изображающих наши объекты, а их взаимное расположение, то условимся считать, что изображения, совмещаемые при помощи движений плоскости или осевых симметрий, эквивалентны.

Теорема 1. *Объем тетраэдра с вершинами в любых четырех точках изучаемого множества равен нулю тогда и только тогда, когда ранг матрицы двойного центрирования B равен 2.*

Доказательство. Допустим ранг B равен хотя бы 3. Тогда по теореме Торгерсона существует изображение в трехмерном пространстве с сохранением всех расстояний. При этом изображение не является двумерным, поскольку иначе ранг матрицы был бы равен двум. Следовательно, можно выбрать четыре точки изображения, не лежащие в одной плоскости. Для них объем соответствующего тетраэдра не равен 0. Противоречие доказывает необходимость.

В обратную сторону. Пусть ранг B равен 2. По теореме Торгерсона существует изображение на двумерной плоскости с сохранением всех расстояний. Все точки лежат в одной плоскости, а значит любые четыре компланарны. То есть объем тетраэдра с вершинами в любых четырех точках равен 0. Теорема доказана. □

Теорема 2. *Пусть дана матрица расстояний D, для которой ранг матрицы B с двойным центрированием равен 2. Тогда конфигурация точек X_{mds}, полученная методом классического многомерного шкалирования, и конфигурация X_{geo}, полученная геометрическим методом, эквиваленты.*

Доказательство. Согласно доказанному выше следствию из теоремы Торгерсона для матрицы B ранга 2, существует конфигурация точек X_{mds} в двумерном пространстве такая что евклидовы расстояния между ними точно равны элементам матрицы D. Из условия $rank(B) = 2$ следует, что объем тетраэдра с вершинами в любых четырех точках изображения равен 0, а значит точки X_{geo} также изображаются без искажения расстояний.

Из университетских курсов геометрии, например, [5], известно, что конфигурация множества точек определяется набором их попарных расстояний однозначно с точностью до движения, т.е. конфигурация X_{mds}, переводится в X_{geo} с помощью симметрий, сдвигов и поворотов, что и завершает доказательство. □

4. Пример

Для сравнения методов возьмем 6 точек трехмерного пространства, лежащие в плоскости, заданной уравнением $2x - y - z + 1 = 0$. Тогда неискаженное их изображение, будет заведомо возможно

Таблица 1

Координаты точек в примере

	x	y	z
A_1	0	0	1
A_2	1	0	3
A_3	2	3	2
A_4	-2	-2	-1
A_5	2	0	5
A_6	1	2	1

По этим данным вычислим матрицу попарных расстояний и выполним оба алгоритма.

Оба метода были реализованы в виде компьютерной программы на языке Python.

Вычислив попарные расстояния для обоих алгоритмов, можно убедиться, что они точно совпадают с исходными. Более того, при визуальном анализе можно заметить, если изображение, полученное геометрическим методом, повернуть по часовой стрелке на 45 градусов, то оно совпадет с изображением многомерного шкалирования.

```
Геометрический метод:
A₁: (0.0, 0.0)
A₂: (2.236, 0.0)
A₃: (1.789, 3.286)
A₄: (-2.683, -2.191)
A₅: (4.472, 0.0)
A₆: (0.447, 2.191)
```

Рисунок 1. Вывод в консоль результатов работы геометрического метода

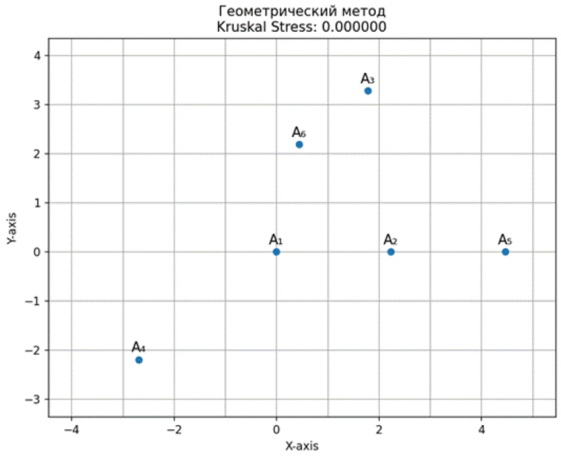

Рисунок 2. Итоговое изображение для геометрического метода

```
Классическое MDS:
A₁: (-1.177, 0.054)
A₂: (0.753, -1.075)
A₃: (2.026, 1.988)
A₄: (-4.6, -0.482)
A₅: (2.683, -2.204)
A₆: (0.315, 1.719)
```

Рисунок 3. Вывод в консоль результата работы классического многомерного шкалирования

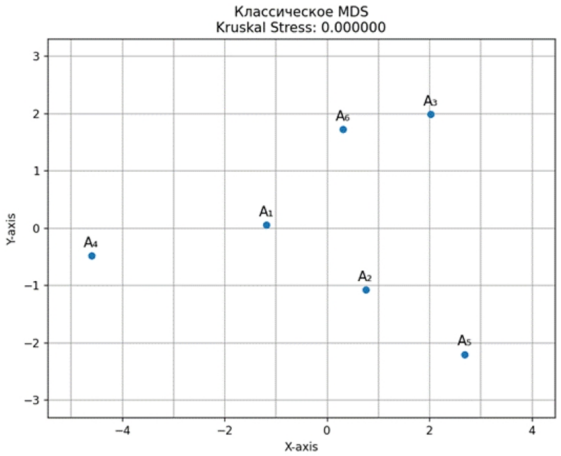

Рисунок 4. Итоговое изображение для классического многомерного шкалирования

5. Заключение

В работе показано, что при наличии возможности неискаженной визуализации сложный матричный аппарат классического многомерного шкалирования и интуитивный геометрический алгоритм приводят к идентичным результатам. Это позволяет использовать геометрический подход как более интерпретируемый инструмент для анализа небольших выборок данных.

Список литературы

1. Young G., Householder A.S. Discussion of a set of points in terms of their mutual distances // Psychometrika. — 1938. — Vol. 3. — P. 19–22.

2. Torgerson W.S. Multidimensional scaling: I. Theory and method // Psychometrika. — 1952. — Vol. 17. — P. 401–419.

3. Ланкастер П., Тисменецкий М. Теория матриц. — М. : Наука, 1988. — 270 с.

4. Калинкин А.А. Неитерационный алгоритм визуализации многомерных данных // Труды семинара по геометрии и математическому моделированию. — 2023. — № 9. — С. 105–111.

5. Беклемишев Д.В. Курс аналитической геометрии и линейной алгебры: Учеб. для вузов. — 11-е, испр. изд. — М. : Физматлит, 2006. — 312 с.