Математические структуры и моделирование
2015. № 4(36). С. 117–122

УДК 519.237.07

МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ НА БАЗЕ МЕТОДА ВЕРЛЕ

В.А. Шовин
научный сотрудник, e-mail: v.shovin@mail.ru

Омский филиал Федерального государственного бюджетного учреждения науки
Института математики им. С.Л. Соболева Сибирского отделения РАН

Аннотация. Предложен новый метод факторизации посредством многомерного шкалирования на базе метода Верле. Результаты факторизации корреляционной матрицы с помощью данного метода находятся в соответствии с классическими методами (метод главных компонент, метод минимальных остатков).

Ключевые слова: метод Верле, многомерное шкалирование, факторный анализ.

Введение

Многомерное шкалирование позволяет в рамках гипотезы о размерности целевого пространства расположить объекты по их взаимным расстояниям таким образом, чтобы восстанавливаемые расстояния между объектами приближались к эмпирическим.

Метод Верле — это итерационный метод вычисления следующего местоположения точки по текущему и прошлому. Этот метод позволяет учитывать дополнительные ограничения, накладываемые на системы точек, например, расстояния между точками. На базе метода Верле предлагается осуществить многомерное шкалирование, тем самым взаимные расстояния между точками будут учтены с наибольшей точностью.

В качестве матрицы взаимных расстояний будет выступать матрица корреляций. С помощью многомерного шкалирования будет осуществлена факторизация корреляционной матрицы, тем самым будет восстановлена факторная структура данных в факторном пространстве. Чтобы получить интерпретабельное решение, предлагается использовать отдельные методы факторного вращения, применённые к восстановленной факторной структуре.

1. Метод Верле

Алгоритм Верле используется для вычисления следующего положения точки по текущему и прошлому:

$$\bar{x}_j^i = \bar{x}_j^{i-1} + \bar{v}_j,$$

$\bar{x}_j^i = \left(x_{j1}^i, x_{j2}^i, \ldots, x_{jm}^i\right)$ — вычисляемые координаты j-ой точки на i-ой итерации,

m — размерность пространства,

$\bar{v}_j = \bar{x}_j^{i-1} - \bar{x}_j^{i-2}$ — вектор скорости j-ой точки.

На систему точек накладываются ограничения.

Некоторые из точек связаны упругими стержнями заданной длины.

Алгоритм работает следующим образом:

1. Вычисляются новые положения точек.
2. Для каждой связи удовлетворяется соответствующее условие.
3. Шаг 2 повторяется s раз.

Например, $s = 16$.

Процедура релаксации связи описывается следующими формулами:

Если связь представлена точками $\bar{a}$ и $\bar{b}$с равновесным расстоянием между ними t, то

$$\bar{a}^i = \bar{a}^{i-1} + \bar{r},$$

$$\bar{b}^i = \bar{b}^{i-1} - \bar{r},$$

$$\bar{r} = f \cdot q \cdot \frac{t - \left|\bar{a}^{i-1} - \bar{b}^{i-1}\right|}{\left|\bar{a}^{i-1} - \bar{b}^{i-1}\right|} \left(\bar{a}^{i-1} - \bar{b}^{i-1}\right),$$

$f = 0.7$ — коэффициент упругости связи,

$q = \dfrac{1}{s}$ — коэффициент, зависящий от числа s повторений шага 2.

2. Многомерное шкалирование

Многомерное шкалирование (МНШ) - это способ наиболее эффективного размещения объектов, приближённо сохраняющий наблюдаемые между ними расстояния. МНШ размещает объекты в пространстве заданной размерности и проверяет, насколько точно полученная конфигурация сохраняет расстояния между объектами. МНШ использует алгоритм минимизации некоторой функции, оценивающей качество получаемых вариантов отображения.

Мерой, наиболее часто используемой для оценки качества подгонки модели (отображения), измеряемого по степени воспроизведения исходной матрицы сходств, является так называемый стресс. Величина стресса φ для текущей конфигурации определяется так:

$$\varphi = \sum_{i=1, j=i+1}^{m} \left(d_{ij} - f\left(\delta_{ij}\right)\right)^2.$$

Здесь d_{ij} — воспроизведённые расстояния в пространстве заданной размерности, а δ_{ij} — исходное расстояние, m — количество объектов. Функция $f(\delta_{ij})$ обозначает неметрическое монотонное преобразование исходных данных (расстояний). МНШ воспроизводит не количественные меры сходств объектов, а лишь их относительный порядок. Чем меньше значение стресса, тем лучше матрица исходных расстояний согласуется с матрицей результирующих расстояний.

3. Главные компоненты и факторная модель

Модель главных компонент описывается следующими формулами

$$\overrightarrow{z_i} = a_{i1}\overrightarrow{p_1} + a_{i2}\overrightarrow{p_2} + \ldots + a_{ig}\overrightarrow{p_g} + d_i\overrightarrow{u_i}$$

m — число переменных,
g — число факторов,
$\overrightarrow{z_i}$ — исходные переменные,
$\overrightarrow{p_i}$ — общие факторы,
$\overrightarrow{u_i}$ — специфичные факторы.

Корреляции между исходными переменными могут быть определены как скалярное произведение нормализованных последних с нулевым средним и единичной дисперсией:

$$r_{ij} = \frac{1}{n+1}\overrightarrow{z_i} \cdot \overrightarrow{z_j},$$

n — размерность исходного пространства переменных.

Коэффициенты корреляций между исходными переменными определяют матрицу корреляций, сходную с матрицей взаимных расстояний метода многомерного шкалирования. Поскольку ближним объектам в факторном пространстве соответствуют большие значения коэффициентов корреляций, то элементы матрицы взаимных расстояний *dij* получаются из соответствующих элементов матрицы корреляций *rij* по формуле:

$$d_{ij} = 1 - |r_{ij}|\,.$$

С помощью метода Верле будет восстановлена факторная структура в рамках гипотезы о размерности g факторного пространства.

Факторная структура

Элементы факторной структуры a_{ij} могут быть определены как коэффициент корреляции между *j*-ой факторной осью и *i*-ой переменной:

$$a_{ij} = \frac{\bar{s}_i \cdot \bar{f}_j}{|\bar{s}_i| \cdot \left|\bar{f}_j\right|},$$

$$\bar{f}_i = (\delta_{i1}, \ldots, \delta_{ig})\,,$$

$\delta_{ij} = \begin{cases} 1, i = j, \\ 0, i \neq j \end{cases}$ — символ Кронекера,

$\bar{s}_j = \bar{h}_j - \bar{h}_c$ — вектор направления j-ой переменной в факторном пространстве.

$\bar{h}_j$ — j-ая переменная в факторном пространстве.

$\bar{h}_c$ — центр масс факторной структуры переменных в факторном пространстве.

4. Численный эксперимент

В качестве исходных параметров были взяты 15 биофизических показателей для 131 лица с артериальной гипертензией начальной стадии:

1) *вес,*
2) *индекс массы тела (ИМТ),*
3) *частота дыхания (ЧД),*
4) *сегментоядерные нейтрофилы (С),*
5) *лимфоциты (Л),*
6) *конечно-систолический размер левого желудочка (КСР),*
7) *конечно-систолический объем левого желудочка (КСО),*
8) *конечно-диастолический размер левого желудочка (КДР),*
9) *конечно-диастолический объем левого желудочка (КДО),*
10) *ударный объем (УО),*
11) *минутный объем сердца (МОС),*
12) *общее периферическое сосудистое сопротивление (ОПСС),*
13) *индекс Хильдебрандта (ИХ),*
14) *фракция выброса левого желудочка (ФВ),*
15) *фракция укорочения левого желудочка (ФУ).*

Программная реализация

Метод Верле был реализован программно с использованием общедоступной JavaScript библиотеки Verlet.js, которая была усовершенствована для многомерного случая. Web-приложение многомерного шкалирования на базе метода Верле доступно по адресу: http://svlaboratory.org/application/multscal — после регистрации нового пользователя. Приложение позволяет визуализировать процесс сходимости метода Верле в заданной плоскости координат (рис. 1).

Результирующая факторная структура для данных артериальной гипертензии представлена в таблице 1.

Факторное решение после факторного вращения по критерию интерпретабельности, предложенное в работе [1], представлено в таблице 2. Данные факторные структуры подтверждаются предыдущими работами [2].

5. Заключение

Метод многомерного шкалирования на базе метода Верле, применённый к корреляционной матрице, является альтернативным методом факторизации.

Таблица 1. Исходное факторное решение (метод Верле)

	*F*1	*F*2	*F*3	*F*4	*F*5
Вес	-0,469	0,442	0,227	-0,667	0,296
ИМТ	-0,336	0,172	0,421	-0,748	0,348
ЧД	-0,258	0,239	0,479	0,68	0,429
С	-0,854	-0,444	0,034	0,032	-0,265
Л	-0,766	-0,584	0,093	0,005	-0,254
КСР	0,914	0,151	-0,27	-0,155	-0,214
КСО	0,931	0,11	-0,251	-0,176	-0,165
КДР	0,604	0,096	-0,783	-0,105	-0,044
КДО	0,667	0,109	-0,727	-0,121	-0,011
УО	0,334	0,024	-0,931	-0,099	0,104
МОС	0,376	0,003	-0,908	0,062	0,175
ОПСС	0,156	0,248	-0,855	0,4	-0,155
ИХ	-0,189	-0,067	0,492	0,592	0,605
ФВ	0,576	-0,123	0,687	-0,211	-0,37
ФУ	0,266	0,085	0,679	0,075	-0,674

Таблица 2. Факторная структура по критерию интерпретабельности (косоугольный случай)

	*F*1	*F*2	*F*3	*F*4	*F*5
Вес	-0,274	-0,251	-0,227	0	**0,899**
ИМТ	-0,352	0	0	-0,198	**0,902**
ЧД	-0,430	-0,203	0,361	**0,765**	-0,238
С	**-0,818**	-0,118	-0,209	-0,412	-0,197
Л	**-0,822**	0	-0,102	-0,515	-0,220
КСР	**0,889**	0,342	-0,061	-0,042	-0,023
КСО	**0,885**	0,353	0	-0,067	0
КДР	**0,940**	-0,210	-0,061	-0,076	-0,039
КДО	**0,963**	-0,163	-0,019	-0,057	-0,009
УО	**0,809**	-0,487	-0,009	-0,110	0
МОС	**0,812**	-0,507	0,105	0	-0,114
ОПСС	**0,631**	-0,531	-0,254	0,281	-0,417
ИХ	-0,453	-0,149	0,653	**0,548**	-0,205
ФВ	0	**0,967**	0,003	-0,219	0,002
ФУ	-0,218	**0,867**	-0,364	0	-0,257

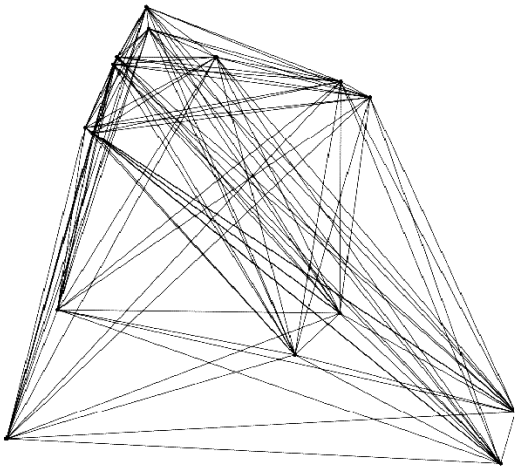

Рис. 1. Визуализация метода Верле в web-приложении

Для 15 биофизических показателей артериальной гипертензии начальной стадии была получена факторная структура на базе данного метода. Полученная факторная структура находится в соответствии с другими методами факторизации.

Литература

1. Шовин В.А., Гольтяпин В.В. Методы вращения факторных структур // Математические структуры и моделирование. 2015. № 2(34). С. 75–83.
2. Гольтяпин В.В., Шовин В.А. Косоугольная факторная модель артериальной гипертензии первой стадии // Вестник Омского университета. 2010. № 4. С. 120–128.

MULTIDIMENSIONAL SCALING BASED METHOD VERLET

V.A. Shovin
Researcher, e-mail: v.shovin@mail.ru

Omsk Branch of the Federal State budget institution Science Institute of Mathematics S.L. Soboleva of Siberian Branch of RAS

Abstract. A new method of factorization through multidimensional scaling based on Verlet method is proposed. The results of the correlation matrix factorization using this method are in accordance with the classical methods (principal component, the minimal residual method).

Keywords: Verlet method, multidimensional scaling, factor analysis.