

Приборостроение, метрология и информационно-измерительные приборы и системы

УДК 681.391:543/545

ЭТАПЫ РЕАЛИЗАЦИИ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ПРИ ОБРАБОТКЕ СИГНАЛОВ АНАЛИТИЧЕСКИХ ПРИБОРОВ

Р.Т. Сайфуллин, С.С. Александров

Самарский государственный технический университет
Россия, 443100, г. Самара, ул. Молодогвардейская, 244

Аннотация. Аналитический прибор позволяет регистрировать данные большим количеством детекторов в течение длительного времени. Актуальным при этом является процесс сжатия и классификации огромного массива полученной информации. Для выделения значимой информации используются методы сжатия данных, основанные на представлении исходных данных через новые переменные меньшей размерности. Одним из способов сжатия данных является метод главных компонент (МГК). В работе произведен анализ основных этапов МГК при обработке сигналов аналитических приборов.

Ключевые слова: измерения, метод сжатия, главные компоненты, коэффициент корреляции.

Современные аналитические приборы могут производить огромное количество измерений. Однако из-за мультиколлинеарности доля полезной информации в таком массиве данных может быть относительно невелика. Для выделения значимой информации используются методы сжатия данных, основанные на представлении исходных данных через новые переменные существенно меньшей размерности, чем число исходных переменных. Сжатие данных позволяет представить полезную информацию в более компактном виде, удобном для визуализации и интерпретации. Одним из основных способов сжатия данных является метод главных компонент (МГК) [1]. МГК является разновидностью мультикорреляционного анализа и основан на обработке корреляционных матриц большой размерности.

При использовании МГК предполагается, что несколько измеряемых переменных сильно коррелируют друг с другом. Это означает, что либо они взаимно определяют друг друга, либо связь между ними обуславливается третьей величиной, которую непосредственно измерить нельзя [2]. Модель главных компонент в большей степени связана с этим предположением и дает возможность по-

Сайфуллин Раухат Талгатович (д.т.н., проф.), профессор кафедры «Информационно-измерительная техника».

Александров Сергей Сергеевич, аспирант.

лучить числовые значения этих третьих величин в виде набора линейно-независимых факторов (ЛНФ), которые описывают и воспроизводят исходную матрицу данных с необходимой точностью.

В общем случае представление данных в виде матрицы «объект – переменная» позволяет решать два вида задач: анализ взаимосвязей между переменными – столбцами матрицы (исследование структуры связей, снижение размерности) и выявление сходства между объектами – строками матрицы (классификация). В обоих случаях целью является «сжатие» информации. Методы снижения размерности «сжимают» матрицу по столбцам, выделяя группы связанных друг с другом переменных. Методы классификации «сжимают» матрицу данных по строкам, объединяя в кластеры схожие между собой объекты. Таким образом, речь идет об агрегировании столбцов или строк исходной матрицы данных.

МГК базируется на геометрической интерпретации пространства переменных и состоит в том, чтобы выделить в многомерном пространстве группы тесно коррелирующих между собой переменных и заменить их интегральными индексами (главными компонентами), которые сохранили бы большую часть исходной информации. Другими словами, МГК позволяет заменить набор из k исходных переменных $x_1, x_2 \dots x_k$ набором из m новых переменных (главных компонент) $y_1, y_2 \dots y_m$, причем $m \ll k$, и сохранить при этом большую часть исходной информации.

Для конструирования главных компонент используется линейная модель в стандартизированных переменных:

$$y_i = \sum_{j=1}^k \alpha_{i,j} z_j, \quad (1)$$

где y_i – главная компонента с номером i ($i = 1, \dots, m$);

z_j – стандартизированная исходная переменная x_j ($z_j = \frac{x_j - \bar{x}_j}{s_j}$);

$\bar{x}_j$ – оценка среднего, ($j = 1, \dots, k$);

$\alpha_{i,j}$ – коэффициент, отражающий вклад переменной z_j в главную компоненту y_i .

Алгоритм МГК состоит из нескольких последовательных этапов.

1. Стандартизация исходных переменных, т. е. переход от исходного пространства переменных $x_1, x_2, \dots, x_k$ к пространству стандартизированных переменных $z_1, z_2, \dots, z_k$ по формуле вычисления z-оценок:

$$z_j = \frac{x_j - \bar{x}_j}{s_j}, \quad (j = 1, \dots, k).$$

Среднее арифметическое стандартизированных переменных равно нулю ($z_j = 0$), дисперсия и стандартное отклонение равны единице ($s_j^2 = s_j = 1$). Следовательно, стандартизированные переменные $z_1, z_2, \dots, z_k$ имеют одинако-

вую информативность, а суммарный объем заключающейся в них информации равен $k \left(\sum_{j=1}^k s_j^2 = k \right)$.

2. Линейное преобразование пространства стандартизированных переменных с целью построения нового ортогонального пространства $y_1, y_2, \dots, y_k$. Линейное преобразование осуществляется по формуле

$$y_i = \sum_{j=1}^k \alpha_{i,j} z_j, \quad (i = 1, \dots, k),$$

где y_i – новая переменная с номером i ($i = 1, \dots, k$); z_j – стандартизированная переменная с номером j ($j = 1, \dots, k$); $\alpha_{i,j}$ – коэффициенты перехода от набора переменных $z_1, z_2, \dots, z_k$ к набору переменных $y_1, y_2, \dots, y_k$.

Более подробно формулу (1) можно представить в виде системы линейных уравнений:

$$\begin{cases} y_1 = \alpha_{1,1}z_1 + \alpha_{1,2}z_2 + \dots + \alpha_{1,k}z_k; \\ y_2 = \alpha_{2,1}z_1 + \alpha_{2,2}z_2 + \dots + \alpha_{2,k}z_k; \\ \dots \\ y_k = \alpha_{k,1}z_1 + \alpha_{k,2}z_2 + \dots + \alpha_{k,k}z_k. \end{cases} \quad (2)$$

Коэффициенты $\alpha_{i,j}$ вычисляются таким образом, чтобы выполнялись следующие условия:

1) дисперсии новых переменных y_i численно равны собственным значениям исходной матрицы корреляций $s^2(y_i) = \lambda_i$. Сумма собственных значений матрицы корреляций $\sum_{i=1}^k \lambda_i = k$, а следовательно, информация, содержащаяся в наборе стандартизированных переменных $z_1, z_2, \dots, z_k$, полностью сохраняется в наборе новых переменных $y_1, y_2, \dots, y_k$;

2) переменные $y_1, y_2, \dots, y_k$ пронумерованы в порядке убывания дисперсий: $s^2(y_1) \geq s^2(y_2) \geq \dots \geq s^2(y_k)$;

3) переменные y_i ($i = 1, \dots, k$) ортогональны, т. е. не коррелируют друг с другом.

Таким образом, получено новое пространство переменных $y_1, y_2, \dots, y_k$, размерность которого совпадает с размерностью исходного пространства. Новое пространство ортогонально, и переменные в нем упорядочены по убыванию дисперсии.

Заметим, что средние арифметические новых переменных равны нулю: $\bar{y}_i = 0$ ($i = 1, \dots, k$), а их суммарная дисперсия $\sum_{i=1}^k s^2(y_i) = k$.

3. Определение числа главных компонент. Суть МГК состоит в том, чтобы сократить размерность пространства переменных посредством «отсечения» некоторого количества наименее информативных переменных с максимальными номерами. Оставшиеся m переменных $y_1, y_2, \dots, y_m$ ($m \ll k$) называются главными компонентами. В процессе выделения главных компонент система уравнений (1) сокращается до m уравнений:

$$y_i = \sum_{j=1}^k a_{i,j} z_j, (i=1, \dots, m).$$

Очевидно, что сумма дисперсий главных компонент меньше, чем сумма дисперсий исходных переменных (k). Долю сохраненной информации можно определить по формуле $I = \frac{\sum_{i=1}^m \lambda_i}{k}$. Соответственно доля утраченной информации

составляет $1 - I = \frac{\sum_{i=m+1}^k \lambda_i}{k}$.

Ключевым моментом выделения главных компонент является определение их количества (m). Эта задача не имеет однозначного решения. Можно использовать следующие критерии для определения числа главных компонент.

Критерий, основанный на собственных числах матрицы корреляции, заключается в том, чтобы ограничить отбор главных компонент теми переменными y_i , которым соответствуют собственные значения $\lambda_i \geq 1$, т. к. их информационная ценность ($s^2(y_i) \geq 1$) заведомо выше информационной ценности отсеченных переменных ($s^2(y_i) < 1$).

Критерий, основанный на доле сохраненной дисперсии, состоит в том, чтобы суммарная дисперсия главных компонент составляла не менее заданной доли исходной суммы дисперсий k . При применении данного критерия рекомендуется использовать накопленные относительные величины собственных значений:

$$\frac{\lambda_1}{k}, \frac{(\lambda_1 + \lambda_2)}{k}, \dots, \frac{(\lambda_1 + \lambda_2 + \dots + \lambda_m)}{k}.$$

Таким образом, все подходы к определению количества главных компонент основаны на дисперсии переменных $y_1, y_2, \dots, y_k$, а она, в свою очередь, определяется формальными характеристиками матрицы корреляций – ее собственными значениями $\lambda_1, \lambda_2, \dots, \lambda_k$.

4. Представление результатов МГК. Результаты МГК принято представлять в виде матрицы коэффициентов линейных преобразований $\alpha_{i,j}$ из формул (1) и (2), которые называются нагрузками на главные компоненты, или факторными нагрузками (табл. 1). Заметим, что нагрузки нумеруются не так, как элементы большинства матриц: первый индекс здесь соответствует номеру столбца таблицы, а второй – номеру строки.

Таблица 1

Полная матрица нагрузок

Переменная	y_1	y_2	...	y_m	...	y_k
z_1	$\alpha_{1,1}$	$\alpha_{2,1}$	...	$\alpha_{m,1}$	...	$\alpha_{k,1}$
z_2	$\alpha_{1,2}$	$\alpha_{2,2}$	...	$\alpha_{m,2}$	...	$\alpha_{k,2}$
...	...	...	...	...	...	
z_k	$\alpha_{1,k}$	$\alpha_{2,k}$	...	$\alpha_{m,k}$	...	$\alpha_{k,k}$

Нагрузки на главные компоненты $\alpha_{i,j}$ ($i = 1, \dots, m; j = 1, \dots, k$) выполняют в МГК три функции:

- 1) используются для определения качества построенной модели;
- 2) служат коэффициентами корреляции между исходными переменными и главными компонентами и в этом значении используются для интерпретации последних;
- 3) используются как коэффициенты линейных уравнений (1) при вычислении значений главных компонент.

Для определения качества модели используются сохраненная дисперсия и общности, вычисляемые как суммы квадратов нагрузок на главные компоненты (табл. 2).

Таблица 2

Показатели качества модели: сохраненная дисперсия и общности

Стандартизированные исходные переменные	Главные компоненты				Общность
	y_1	y_2	...	y_m	
z_1	$\alpha_{1,1}$	$\alpha_{2,1}$	...	$\alpha_{m,1}$	h_1^2
z_2	$\alpha_{1,2}$	$\alpha_{2,2}$	...	$\alpha_{m,2}$	h_2^2
...	...	...	...	...	...
z_k	$\alpha_{1,k}$	$\alpha_{2,k}$	...	$\alpha_{m,k}$	h_k^2
Дисперсия	$s_1^2 = \lambda_1$	$s_2^2 = \lambda_2$	...	$s_m^2 = \lambda_m$	$\sum_{i=1}^m s_i^2 = \sum_{j=1}^k h_j^2$
Доля сохраненной дисперсии	$\frac{s_1^2}{k}$	$\frac{s_2^2}{k}$	...	$\frac{s_m^2}{k}$	—
Накопленная доля сохраненной дисперсии	$\frac{s_1^2}{k}$	$\frac{s_1^2 + s_2^2}{k}$	...	$\frac{\sum_{i=1}^m s_i^2}{k}$	—

Сумма квадратов нагрузок по столбцу с номером i равна дисперсии главной компоненты y_i и, соответственно, собственному значению λ_i матрицы корреляций: $s_i^2 = \sum_{j=1}^k \alpha_{i,j}^2 = \lambda_i$ ($i = 1, \dots, m$). После деления дисперсии на общий объем информации в исходном наборе переменных k получаем долю дисперсии, сохраненной в соответствующей главной компоненте $\frac{s_i^2}{k}$. В последней строке табл. 2 указана накопленная доля сохраненной дисперсии для главных компонент с соответствующими номерами.

Сумма квадратов нагрузок по строке с номером j $h_j^2 = \sum_{i=1}^m \alpha_{i,j}^2$ называется общностью исходной переменной x_j ($j = 1, \dots, k$) и показывает долю ее дисперсии, сохраненной в m главных компонентах.

5. Интерпретация главных компонент. Матрица нагрузок (см. табл. 1) используется также для интерпретации главных компонент. Поскольку нагрузки являются коэффициентами корреляции между главными компонентами и исходными переменными, для интерпретации используются переменные, имеющие максимальные по абсолютной величине нагрузки.

При интерпретации нагрузок различают две основные ситуации:

1) если все значительные по абсолютной величине нагрузки имеют одинаковые знаки, главная компонента называется главной компонентой размера; она показывает степень выраженности у объектов одной латентной характеристики, которую и следует интерпретировать;

2) если значительные по абсолютной величине нагрузки имеют разные знаки, то говорят о главной компоненте формы, которая дифференцирует объекты из выборки в соответствии с наличием у них двух свойств, в некоторой степени противоположных друг другу.

Для облегчения восприятия факторной структуры может применяться графическое представление нагрузок. Для этого в пространстве двух, максимум трех главных компонент стандартизированные исходные переменные $x_1, x_2, \dots, x_k$ изображаются в виде точек, в качестве координат используются соответствующие значения нагрузок.

6. Вычисление главных компонент. Наиболее распространенным методом вычисления главных компонент является регрессионное шкалирование, которое заключается в том, что для каждого объекта из выборки в уравнение (1) подставляются стандартизированные значения исходных переменных. Вычисление значений главных компонент фактически приводит к окончательному переходу из исходного пространства переменных в пространство главных компонент [3].

Одним из возможных преимуществ МГК является возможность представить сложные данные физического эксперимента в более простом виде, так чтобы исследователь смог увидеть результаты эксперимента в простой графической интерпретации. В общем случае, если структура данных такова, что в них имеются какие-то группы или кластеры, это видно при исследовании проекций, построенных для первых нескольких компонент. Поэтому при исследовании данных ме-

тодом МГК особое внимание уделяется графикам нагрузок. Они несут в себе информацию для понимания того, как устроены данные.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Померанцев А.Л. Метод главных компонент [Электронный ресурс] // Российское хеометрическое общество. – 2008. – Режим доступа: <http://rscs.chph.ras.ru/>
2. Гаврилина В.А., Сычев С.Н. Исходные гипотезы для распознавания многокомпонентных физико-химических систем комбинацией «высокоэффективная жидкостная хроматография – метод главных компонент» // Сорбционные хроматографические процессы. – 2012. – Т. 12. – Вып. 5. – С. 798–805.
3. Терещенко О.В., Курилович Н.В., Князева Е.И. Многомерный статистический анализ данных в социальных науках: Учеб. пособие. – Минск: БГУ, 2012. – 239 с.

Статья поступила в редакцию 20 февраля 2018 г.

STAGES OF REALIZATION OF THE PRINCIPAL COMPONENT METHOD IN PROCESSING OF SIGNALS FROM ANALYTICAL INSTRUMENTS

R.T. Saifullin, S.S. Aleksandrov

Samara State Technical University
244, Molodogvardeyskaya st., Samara, 443100, Russian Federation

Abstract. *The analytical device allows to register data with a large number of detectors for a long time. The process of compression and classification of a huge array of information received is relevant. Data compression methods based on the representation of the source data through new variables of smaller dimension are used to highlight significant information. One method of data compression is principal component method (PCM). In work presented the main stages of the PCM in the processing of signals from analytical instruments.*

Keywords: *measurements, compression method, principal component method, correlation coefficient.*

*Rauhat T. Saifullin (Dr. Sci. (Techn.)), Professor.
Sergey S. Aleksandrov, Postgraduate Student.*