Е.Г. Галицкая, Е.Б. Галицкий
(*Москва*)

КЛАСТЕРЫ НА ФАКТОРАХ: КАК ИЗБЕЖАТЬ РАСПРОСТРАНЕННЫХ ОШИБОК?

В статье анализируется, почему результаты применения кластерного анализа в факторном пространстве бывают неадекватны структуре экспериментального материала. Предлагается методика, позволяющая избежать указанных искажений.

Ключевые слова: факторный анализ, кластерный анализ, адекватность применения метода, структура экспериментальных данных.

Кластерный анализ на практике нередко применяют в пространстве не многочисленных, как правило, исходных переменных, а нескольких обобщенных показателей. Когда среди исходных показателей есть неметрические (номинальные или порядковые), пока не сформированы метрические обобщенные показатели, применение метода k-means, а при большинстве мер связи – и методов иерархической классификации, просто невозможно[1]. Но и когда все исходные переменные – метрические, то переход к пространству обобщенных переменных – факторов – нередко оказывается

Елена Геннадьевна Галицкая – ведущий специалист Фонда «Общественное мнение», доцент Государственного университета – Высшая школа экономики.
Ефим Борисович Галицкий – кандидат экономических наук, ведущий специалист Фонда «Общественное мнение», доцент Государственного университета – Высшая школа экономики.

[1] Для формирования обобщенных показателей в таких случаях используются такие, например, методы, как анализ гомогенности, начиная с 13-й версии программного пакета *SPSS* называемый нелинейным анализом главных компонент.

плодотворным: за счет концентрации внимания на главных, наиболее типичных различиях в исследуемом материале, он позволяет получить наглядные, хорошо интерпретируемые результаты[1].

Опыт, однако, показывает, что бездумно применять такой подход нельзя. Иногда кластерный анализ в факторном пространстве дает внешне правдоподобные, но абсолютно бессмысленные результаты. Такой эффект, в частности, был блестяще продемонстрирован А.О. Крыштановским[2]. Очень важно разобраться, что именно в построенном им тестовом примере привело к такому итогу, и как на практике избежать получения ошибочных результатов классификации. Ведь пока мы не понимаем, в чем внутренняя причина бессмыслицы, мы рискуем получать ее вновь и вновь! В данной статье обсуждаются такого рода вопросы применительно к случаю метрических исходных переменных, когда обобщенные переменные формируются с помощью классического метода главных компонент.

Посмотрим, как организован разработанный А.О. Крыштановским тестовый материал. Таблица данных содержит 500 строк и 16 столбцов: A, B1, B2, …, B15. В столбце A 250 единиц и 250 двоек, и он используется для расчета остальных столбцов таблицы, которые затем служат исходными данными для анализа. Идея этого расчета в том, что столбцы B1, B2, …, B15 рассчитываются при A=1 – по одному правилу, а при A=2 – по другому, и при этом

[1] Учитывая плодотворность перехода к пространству обобщенных переменных, программный пакет *SPSS* (начиная с 11-й версии), наряду с методом k-means и иерархическим кластерным анализом, содержит двухшаговый метод кластерного анализа, проводящий факторизацию пространства, а затем кластеризацию с подбором оптимального в определенном смысле числа кластеров.

[2] *Крыштановский А.О.* «Кластеры на факторах» – об одном распространенном заблуждении // Социология: методология, методы, математические модели. 2005. № 21. С. 172–187. (Второго августа 2005 г. пришла весть о безвременной кончине нашего дорогого коллеги А.О. Крыштановского. Было бы несправедливо по отношению к памяти Александра Олеговича прервать начатое им обсуждение столь важной темы.)

используется генератор псевдослучайных нормально распределенных чисел. Правило расчета описывается приводимым ниже командным файлом формата SPSS, основные команды которого приведены в статье А.О. Крыштановского.

```
IF (A=1) B1=10*NORMAL(1).
IF (A=2) B1=20+10*NORMAL(1).
COMPUTE B2=0.
…
COMPUTE B15=0.
EXECUTE.
DO REPEAT R=B2 TO B15.
        IF (A=1) R=B1+20*NORMAL(1).
        IF (A=2) R=B1+20*NORMAL(1)+10.
END REPEAT.
EXECUTE.
```

Чтобы легче было представить себе, каких результатов можно было бы ожидать от анализа данных в столбцах B1, B2, …, B15, рассмотрим сначала простейший случай, когда случайных колебаний нет. Получаем следующие правила расчета:
при A=1: B1=0, B2=0, …, B15=0;
при A=2: B1=20, B2=30, …, B15=30.

Данные в столбцах B2, B3, …, B15 в точности совпадают между собой и линейно связаны со столбцом B1, причем коэффициент пропорциональности равен 1,5. Для иллюстрации этого факта покажем, как выглядят экспериментальные точки в плоскости B1-B2 (рис. 1): одни 250 точек проецируются в точку с координатами (0, 0), а другие – с координатами (20, 30).

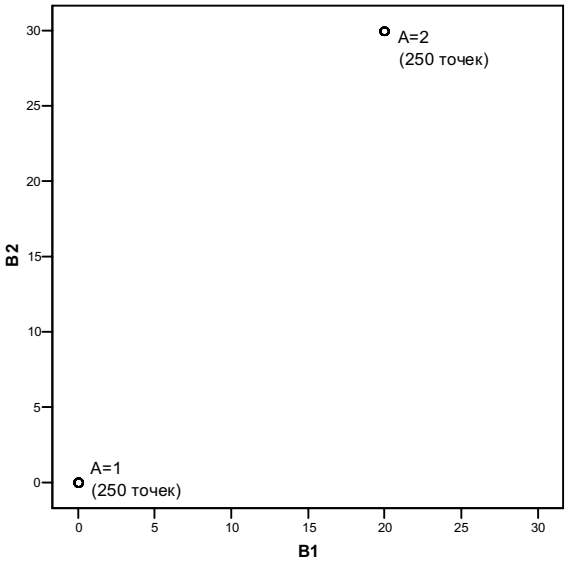

Рис. 1. Расположение экспериментальных точек при отсутствии случайных колебаний

Точно так же выглядят экспериментальные точки в осях B1-B3, B1-B4, …, B1-B15. Очевидно, что при таких данных матрица корреляции между столбцами B1, B2, …, B15 состоит из одних единиц и является вырожденной. Тем не менее, программа факторного анализа из пакета SPSS успешно справляется со своей задачей и строит ровно один фактор, который, естественно, объясняет все 100% дисперсии исходного материала (табл. 1, три правых столбца).

Справляется со своей простейшей в данном случае задачей и кластерный анализ на факторе (алгоритм K-средних): итоговое разбиение строк на два класса в точности совпадает с тестовым столбцом A.

Итак, проблемы в совместной работе этих алгоритмов могут возникать, когда данные содержат случайную помеху. Приведем распределение точек в плоскости B1-B2 для этого случая (рис. 2).

Таблица 1

РАСПРЕДЕЛЕНИЕ МЕЖДУ ФАКТОРАМИ ДИСПЕРСИИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ ПРИ ОТСУТСТВИИ СЛУЧАЙНЫХ ОТКЛОНЕНИЙ

Номер фактора (Components)	Первоначальные собственные числа (Initial Eigenvalues)			Суммы квадратов факторных нагрузок (Extraction Sums of Squared Loadings)		
	Всего (Total)	% дисперсии (% of Variance)	Нарастающим итогом, % (Cumulative %)	Всего (Total)	% дисперсии (% of Variance)	Нарастающим итогом, % (Cumulative %)
1	15,0	100,0	100,0	15,0	100,0	100,0
2	0,0	0,0	100,0			
…	…	…	…			
15	0,0	0,0	100,0			

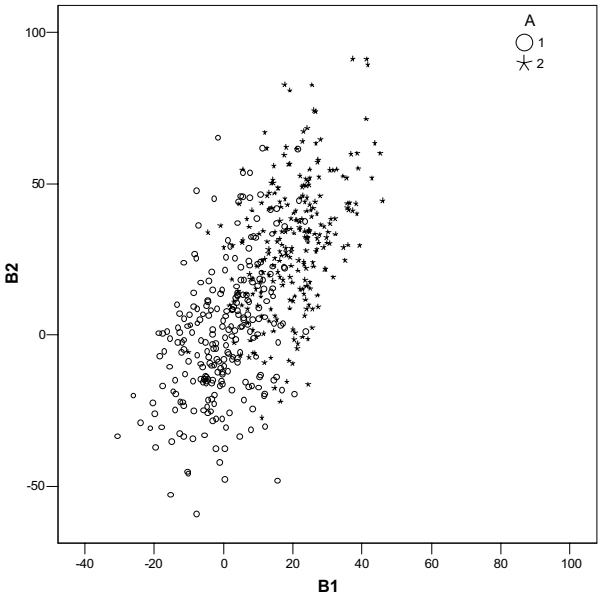

Рис. 2. Расположение экспериментальных точек при наличии случайных колебаний в данных

Белыми кружками на этом рисунке показаны точки, соответствующие строкам с A=1, черными точками – с A=2. Из рисунка видно, что значительная часть наблюдений в результате случайных отклонений «перемешивается», отклоняется от «своего» центра настолько, что оказывается ближе к «чужому». Поэтому результаты работы кластерного анализа не могут не отличаться от столбца A. (Кластерный анализ не обладает аппаратурой распознавания «свой-чужой». В частности, примененный нами метод K-means предназначен для разнесения объектов по классам, исходя из их близости к центру.) И действительно, при кластерном анализе на всех столбцах таблицы данных примерно 8% строк попадают в недиагональные клетки таблицы сопряженности, т.е. классифицируются вместе не со «своими», а с «чужими» строками (табл. 2).

Таблица 2

КРОСС-ТАБУЛЯЦИЯ СТОЛБЦА А И РЕЗУЛЬТАТОВ КЛАСТЕРНОГО АНАЛИЗА НА ВСЕХ СТОЛБЦАХ ИСХОДНОЙ МАТРИЦЫ ДАННЫХ, % по таблице

		Значение в столбце А		Всего (Total)
		1	2	
Номер кластера на столбцах B1, …, B15	1	45,2	3,2	48,4
	2	4,8	46,8	51,6
Всего (Total)		50,0	50,0	100,0

Поскольку, как мы видели, это не ошибки классификации, а результат больших случайных отклонений в экспериментальных данных, будем считать эталоном классификации не столбец А, а результаты работы кластерного анализа на всех столбцах таблицы. Именно с ними мы будем сравнивать классификацию на факторах.

Вернемся к рис. 2. На нем экспериментальные точки образуют размытый эллипс, главная ось которого соединяет показанные на рис. 1 точки (0, 0) и (20, 30), т.е. лежит на линии $B2 = 1{,}5 \cdot B1$. Факторный анализ эту ось эллипса рассеяния легко распознает (рис. 3 и табл. 3). Первый фактор, проходящий через эту ось, объясняет 51,5% дисперсии материала (его собственное число равно 7,7), а на каждый из последующих (перпендикулярных к этой оси) факторов приходится всего от 4,8% до 0,6% дисперсии материала (собственные числа 0,7 и ниже).

Три правых столбца табл. 3 показывают, что если не менять рекомендуемых SPSS установок, будет отобран для дальнейшего анализа только один фактор. О сути этих установок необходимо сделать техническое пояснение.

Как известно, процедура факторного анализа (метод главных компонент) начинает работу с того, что центрирует и нормирует каждый столбец исходных данных. После центрирования и нормирования дисперсия каждого столбца становится, естественно,

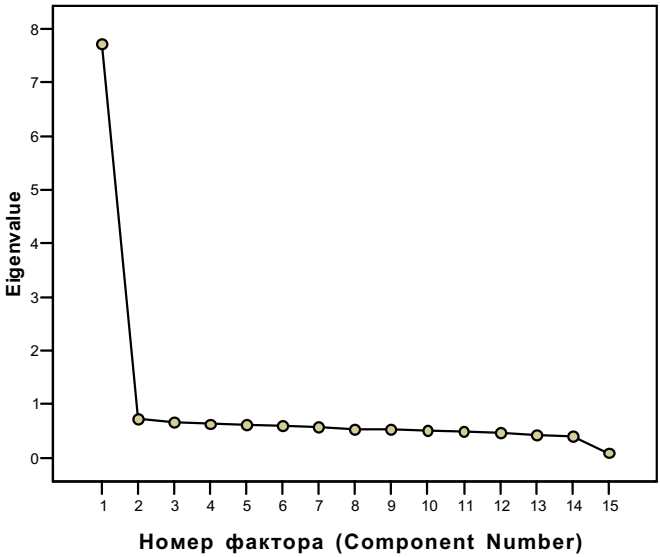

Рис. 3. График зависимости дисперсии, объясняемой фактором, от номера фактора (Screen Plot или «каменистая осыпь»)

равной единице, а дисперсия экспериментального материала в целом[1] – сумме дисперсий столбцов, т.е. числу столбцов.

[1] Имеется в виду дисперсия многомерной (векторной) случайной величины. Как известно, дисперсия векторной случайной величины $X = \{x_1, x_2, ...x_k\}$ в k-мерном пространстве равна сумме дисперсий скалярных случайных величин, которые служат ее координатами. Действительно, имеем:

$$D_X = \frac{1}{N} \cdot \sum_{i=1}^{N} \sum_{j=1}^{k} (x_{i,j} - m_j)^2 = \sum_{j=1}^{k} \frac{1}{N} \cdot \sum_{i=1}^{N} (x_{i,j} - \bar{x}_j)^2 = \sum_{j=1}^{k} D_{x_j},$$

где D_x – дисперсия многомерной случайной величины X, $\bar{x}_j$ и D_{x_j} – соответственно среднее значение и дисперсия одномерной случайной величины x_j, N – число наблюдений.

Таблица 3

РАСПРЕДЕЛЕНИЕ МЕЖДУ ФАКТОРАМИ ДИСПЕРСИИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ ПРИ НАЛИЧИИ СЛУЧАЙНЫХ КОЛЕБАНИЙ В ДАННЫХ

Номер фактора (Components)	Первоначальные собственные числа (Initial Eigenvalues)			Суммы квадратов факторных нагрузок (Extraction Sums of Squared Loadings)		
	Всего (Total)	% дисперсии (% of Variance)	Нарастающим итогом, % (Cumulative %)	Всего (Total)	% дисперсии (% of Variance)	Нарастающим итогом, % (Cumulative %)
1	7,7	51,5	51,5	7,7	51,5	51,5
2	0,7	4,8	56,4			
3	0,7	4,4	60,8			
4	0,6	4,2	65,0			
5	0,6	4,1	69,1			
6	0,6	3,9	73,0			
7	0,6	3,8	76,8			
8	0,5	3,6	80,5			
9	0,5	3,5	84,0			
10	0,5	3,4	87,4			
11	0,5	3,2	90,6			
12	0,5	3,1	93,7			
13	0,4	2,9	96,6			
14	0,4	2,7	99,4			
15	0,1	0,6	100,0			

Факторы пропорциональны собственным векторам матрицы корреляции между столбцами таблицы данных. Технически при желании можно построить ровно столько факторов, сколько столбцов в таблице данных (см. три левых столбца табл. 1). Суммарная дисперсия, объясненная этими факторами, равная сумме собственных чисел, составит многомерную дисперсию экспериментального материала (в нашем примере – 15). Но распределяется эта дисперсия материала между факторами уже не поровну, как между центрированными и нормированными столбцами таблицы исходных данных. Например, в нашем случае дисперсия, объясненная первым фактором, на порядок больше дисперсии, объясненной вторым.

Графически это означает (см. рис. 2), что длина главной оси эллипса рассеяния, соответствующая первому фактору, относится к длине перпендикулярной ей оси, соответствующей второму фактору, как корень квадратный из отношения их собственных чисел, т.е. в следующее число раз:

$$\sqrt{\frac{\lambda_1}{\lambda_2}} = \sqrt{\frac{7,7}{0,7}} = 3,3.$$

Другими словами, эллипс рассеяния экспериментальных данных в нашем случае сильно вытянут. Содержательно причина этой вытянутости эллипса, т.е. столь большого разрыва в дисперсии, объясненной первым и вторым, а также остальными факторами, ясна: метод способен отличить действительные закономерности (в данном случае характеризуемые первым фактором) от случайных колебаний, подстроиться к которым пытаются другие факторы. Если же (как в нашем случае) других закономерностей нет, остаточная дисперсия распределяется между прочими факторами практически поровну.

Исходя из этой логики, по умолчанию процедура факторного анализа отбирает для дальнейшего использования лишь факторы, каждый из которых объясняет больше дисперсии, чем один

столбец исходной матрицы данных после его центрирования и нормирования, т.е. факторы с собственным числом, превышающим единицу. Следуя этому критерию, в нашем случае надо ограничиться одним фактором.

В результате кластерного анализа (процедура К-средних) на оси первого фактора лишь 0,8% от числа всех строк таблицы (4 строки из 500) классифицируются иначе, чем на исходных данных (табл. 4).

Таблица 4

КРОСС-ТАБУЛЯЦИЯ РЕЗУЛЬТАТОВ КЛАСТЕРНОГО АНАЛИЗА НА ОДНОМ ФАКТОРЕ И НА ВСЕХ СТОЛБЦАХ ИСХОДНОЙ ТАБЛИЦЫ ДАННЫХ, % по таблице

		Эталонная классификация (на столбцах B1, …, B15)		Всего (Total)
		1	2	
Кластеры на одном факторе	1	48,4	0,8	49,2
	2	0,0	50,8	50,8
Всего (Total)		48,4	51,6	100,0

Итак, кластерный анализ на одном факторе дает практически тот же результат, что и на всех 15 исходных столбцах таблицы данных.

Как показал А.О. Крыштановский, ситуация может кардинально измениться, если выбрать для анализа не один, а больше факторов. При этом часто исходят из того, что большее число факторов объяснит более высокую долю дисперсии, чем один. Так, в статье А.О. Крыштановского было выбрано четыре фактора, поскольку они объясняют в совокупности 65,0%, а не 51,5% дисперсии[1] (см. табл. 3). В таком случае кластерный анализ действительно

[1] В силу использования генератора случайных чисел наши расчеты несколько отличаются от описанных в статье А.О. Крыштановского. Например, в ней четыре фактора объясняют не 65%, как у нас, а около 67% дисперсии. Однако эти различия содержательно ничего не меняют.

дает очень плохой результат: неправильно классифицируется 125 строк, т.е. ровно четверть (см. табл. 5).

Таблица 5

КРОСС-ТАБУЛЯЦИЯ РЕЗУЛЬТАТОВ КЛАСТЕРНОГО АНАЛИЗА НА ЧЕТЫРЕХ ФАКТОРАХ И НА ВСЕХ СТОЛБЦАХ ИСХОДНОЙ ТАБЛИЦЫ ДАННЫХ, % по таблице

		Эталонная классификация (на столбцах B1, …, B15)		Всего (Total)
		1	2	
Кластеры на четырех факторах	1	37,2	13,8	51,0
	2	11,2	37,8	49,0
Всего (Total)		48,4	51,6	100,0

Причину неправильной работы этого метода иллюстрирует рис. 4.

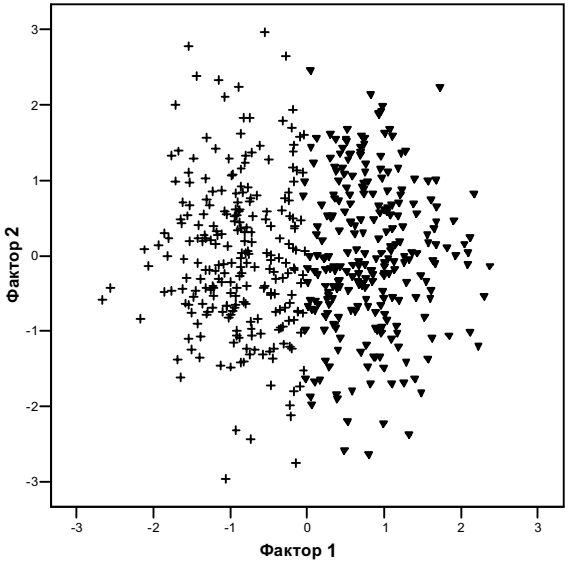

Рис. 4. Расположение экспериментальных точек в осях первых двух факторов

Из приведенного выше рисунка видно, решения какой задачи мы ожидали от кластерного анализа и почему этого сделать не удалось. Крестиками на нем показаны строки, которые, согласно эталонной классификации на исходных столбцах таблицы, относятся к первому кластеру, а черными треугольниками – ко второму. Очевидно, что действительно единственным существенным для распознавания фактором является первый: проекции всех точек второго эталонного кластера на его оси располагаются правее, чем точка первого эталонного кластера. В пространстве же двух факторов эталонные кластеры представляют собою два полушара, расположенных вплотную друг к другу. Однако кластерный анализ (метод К-средних) нацелен на выявление в пространстве не полушаров, а форм, близких к шарообразной.

Итак, неудовлетворительный результат кластерного анализа вполне закономерен. Он является результатом двух обстоятельств: во-первых, выбора излишнего числа факторов и, во-вторых, использования этих факторов без всяких преобразований.

Поясним последнюю мысль. Сравним рис. 2 и рис. 4. На рис. 2 данные представляют собою вытянутый эллипсоид, который кластерный анализ с легкостью «разрезает» на два «почти шара». А на рис. 4 эллипсоид модифицирован практически в шар. Мы считаем правильным разделение точек, изображенных на этом рисунке, вертикальной чертой. Но с точки зрения кластерного анализа любые разделения этого шара на две половинки практически равноправны. Например, если бы генератор случайных чисел сработал немного иначе, то более предпочтительным в смысле компактного расположения точек вполне могло бы оказаться, например, разделение по горизонтали. Тогда ошибок в классификации было бы гораздо больше, чем четверть.

Итак, кластерный анализ не оправдывает наши ожидания из-за «сплющивания» первоначального эллипсоида данных по главной оси. Действительно, при переходе от рис. 2 к рис. 4 мы забыли, что факторы объясняют совершенно разные доли дисперсии

материала, что главная ось первоначального эллипсоида рассеяния, как было показано выше, в 3,3 раза длиннее второй. Другими словами, *различие на определенную величину по оси первого фактора более чем втрое важнее при классификации, чем такое же различие по второму*. Эта информация отсутствует на рис. 4, там факторы полностью равноправны, дисперсия каждого из них равна единице.

Для выхода из этой ситуации мы предлагаем перед выполнением кластерного анализа вернуть каждому фактору ту дисперсию, которую он объясняет в исходном материале[1]. Для этого *каждый фактор умножается на корень квадратный из соответствующего ему собственного числа*. Это преобразование мы используем в нашей практической работе уже много лет, и каждый раз его применение оказывается весьма плодотворным.

Посмотрим, как преобразится рис. 4 после такого преобразования (рис. 5).

Мы видим, что перед нами практически точная, только повернутая, копия первоначального эллипсоида рассеяния, изображенного на рис. 2[2]. Поэтому и результаты кластерного анализа на четырех преобразованных таким образом факторах исключительно точны (табл. 6).

[1] Для простоты мы не обсуждаем здесь подробно, что имеется в виду исходный материал, несколько преобразованный путем стандартизации (центрирования и нормирования) исходных переменных.

[2] Операция стандартизации, о которой говорилось в предыдущей сноске, в данном случае ничего существенного не меняет. В частности, эллипсоид рассеяния в пространстве первых двух осей после стандартизации этих переменных выглядит практически так же, как на рис. 2, только его главная ось лежит на биссектрисе координатного угла.

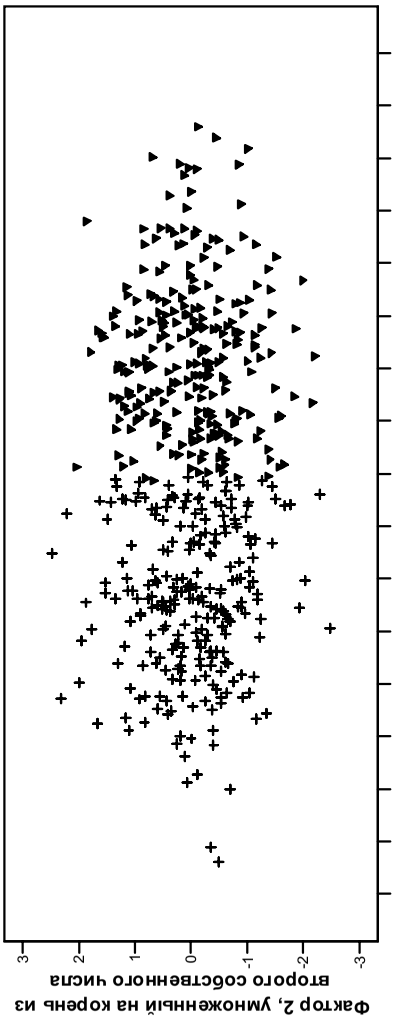

Рис. 5. Расположение экспериментальных точек в осях первых двух преобразованных факторов

Таблица 6

КРОСС-ТАБУЛЯЦИЯ РЕЗУЛЬТАТОВ КЛАСТЕРНОГО АНАЛИЗА НА ЧЕТЫРЕХ ФАКТОРАХ И НА ВСЕХ СТОЛБЦАХ ИСХОДНОЙ ТАБЛИЦЫ ДАННЫХ, % по таблице

		Эталонная классификация (на столбцах B1, …, B15)		Всего (Total)
		1	2	
Кластеры на четырех преобразованных факторах	1	48,4	0,6	49,0
	2	0,0	51,0	51,0
Всего (Total)		48,4	51,6	100,0

Приведенные выше результаты классификации в пространстве преобразованных факторов практически неотличимы от эталонной классификации, полученной на всем наборе исходных столбцов. Более того, различия между ними еще меньше, чем при почти безошибочной классификации на оси одного лишь первого фактора (см. табл. 4): теперь даже не четыре, а лишь три строчки из пятисот классифицированы иначе, чем в эталонной классификации.

* * *

Мы рассмотрели причины, способные привести к серьезным искажениям при выполнении кластерного анализа в пространстве факторов, построенных методом главных компонент. Таких причин две: выбор слишком большого числа факторов и искажение пропорций экспериментального материала после перехода от исходных показателей к факторам. Первая из этих проблем решается путем отбрасывания факторов с собственными числами, меньшими единицы. (Полезен и эмпирический критерий «каменистой осыпи», который в данной статье не обсуждается.) Что же касается второй проблемы, то нами предложена процедура, позволяющая ее полностью преодолеть путем умножения каждого фактора на корень квадратный из соответствующего этому фактору собственного числа.

С учетом сказанного, кластерный анализ на факторах выполнять можно, но при этом необходимо глубоко чувствовать суть применяемых процедур и принимать специальные меры, чтобы избежать опасностей, блестяще продемонстрированных нашим коллегой А.О. Крыштановским в его последней статье.