
А.О. Крыштановский
(Москва)

«КЛАСТЕРЫ НА ФАКТОРАХ» – ОБ ОДНОМ РАСПРОСТРАНЕННОМ ЗАБЛУЖДЕНИИ

Статья посвящена обоснованию некорректности проведения классификации объектов на данных, полученных после факторизации переменных. Эмпирическое обоснование проводится на тестовой матрице данных, сформированной с помощью датчика случайных чисел. В матрице представлены две различные модельные группы респондентов, характеризующиеся 15 переменными.

Ключевые слова: метод факторного анализа, метод главных компонент, кластерный анализ, датчик случайных чисел.

Задача построения классификации единиц исследования является весьма распространенной как в социологических, так и в маркетинговых исследованиях. Получение *однородных групп* объектов (респондентов), т.е. таких групп, которые приблизительно одинаково ведут себя в одинаковых ситуациях (одинаково отвечают на вопросы анкеты) – типичная задача сегментирования.

Определенной проблемой при этом является то, что количество параметров, по которым требуется достижение однородности, во многих случаях весьма велико (нередко – несколько десятков). В этой ситуации непосредственная классификация объектов (как правило, с использованием методов кластерного анализа) приводит к плохо интерпретируемым результатам. Действительно,

Александр Олегович Крыштановский – кандидат философских наук, заведующий кафедрой методов сбора и анализа социологической информации Государственного университета – Высшая школа экономики, декан факультета социологии ГУ-ВШЭ.

кластерный анализ методом *K*-средних (без задания содержательно осмысленных центров кластеров) в качестве исходных точек выбирает максимально далеко отстоящие друг от друга точки, которые на практике часто действительно трудно интерпретируемы. Далее, весь массив разделяется на однородные группы с точки зрения близости к этим «непонятым» объектам. Нет ничего удивительного, что результат становится мало вразумительным.

Распространенным подходом в данной ситуации считается двухстадийный метод, когда на первом этапе к исходным признакам применяется факторный анализ с целью получения некоторых латентных показателей (факторов), объединяющих в некоторые группы (факторы) сами признаки. На втором шаге используют кластерный анализ для получения некоторых групп, однородных в смысле средних величин индивидуальных значений построенных факторов.

На первый взгляд такой подход представляется вполне логичным и естественным. Действительно, в данном случае мы проводим кластеризацию небольшого количества исходных признаков, при котором специфическое поведение даже одного из этих признаков может вызвать сильное смещение результирующей кластеризации, а классифицируем объекты по 3-4 переменным (факторам), каждая из которых при этом имеет более или менее вразумительную интерпретацию.

Данный подход, по нашим наблюдениям, достаточно широко используется в практике как социологических, так и маркетинговых исследований. Такой путь рекомендуется и в достаточно широко распространенной книге А. Бююля и П. Цефеля¹. К сожалению,

¹ Бююль А., Цефель П. SPSS: искусство обработки информации. М.; СПб.; Киев: DiaSoft, 2002. С. 394–398. В данной главе факторный анализ назван почему-то «факториальным», но это, по всей видимости, редакционные погрешности. Отметим, что такой же подход пропагандируется авторами и в разделе, в котором обсуждается кластерный анализ методом *K*-средних (с. 404–409).

внешняя логичность такого подхода никак не учитывает базовых положений метода факторного анализа, которые, как нам представляется, приводят к тому, что из такого рода классификаций хоть сколь-нибудь обоснованных выводов получиться не может.

Для иллюстрации высказанных соображений был проведен описанный ниже эксперимент.

Массив данных

С помощью датчика случайных чисел был создан тестовый массив из 500 объектов, содержащий ответы 2 групп респондентов на 15 вопросов. Файл синтаксиса SPSS по созданию массива приведен ниже.

```
IF (A=1) B1=10*NORMAL (1).
```

```
IF (A=2) B1=20+10*NORMAL (1).
```

*** A – переменная, определяющая принадлежность объекта к одной из 2 групп.**

```
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.
```

Средние значения всех переменных в 2 группах достаточно сильно различаются между собой (табл. 1), и, следовательно, можно считать, что эти 2 группы представляют собой разные совокупности объектов, что, по логике исследования, должно обнаружиться с помощью методов классификации.

Табл. 2 демонстрирует значения коэффициентов корреляции для созданных 15 переменных.

Таблица 1
СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ МОДЕЛЬНЫХ
ПЕРЕМЕННЫХ В 2 ГРУППАХ

Переменная	Номер группы	Среднее значение	Стандартное отклонение
B1	1	-0,88	10,38
	2	19,97	10,03
B2	1	-1,64	23,35
	2	31,48	19,86
B3	1	-0,95	23,23
	2	28,36	21,95
B4	1	-1,32	23,10
	2	29,78	22,32
B5	1	-0,90	23,99
	2	33,07	22,66
B6	1	-0,49	22,71
	2	31,70	22,15
B7	1	-0,12	22,52
	2	30,25	22,44
B8	1	-1,14	21,92
	2	31,27	23,26
B9	1	-0,75	21,79
	2	31,00	22,94
B10	1	0,16	21,77
	2	29,09	21,88
B11	1	-2,24	21,96
	2	31,29	23,70
B12	1	-0,82	23,16
	2	28,06	22,48
B13	1	0,40	24,70
	2	29,43	20,11
B14	1	-0,42	21,21
	2	31,95	22,41
B15	1	-1,99	22,15
	2	29,62	21,64

* Средние значения всех переменных в 2 группах различаются с вероятностью $P > 0,99$.

**МАТРИЦА КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ ПИРСОНА
ДЛЯ 15 МОДЕЛЬНЫХ ПЕРЕМЕННЫХ**

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15
B1	1	0,688	0,662	0,700	0,689	0,660	0,661	0,672	0,661	0,683	0,690	0,689	0,699	0,678	0,651
B2	0,688	1	0,479	0,478	0,509	0,495	0,495	0,461	0,434	0,521	0,465	0,481	0,496	0,512	0,481
B3	0,662	0,479	1	0,452	0,464	0,500	0,466	0,441	0,428	0,442	0,437	0,499	0,522	0,459	0,433
B4	0,700	0,478	0,452	1	0,507	0,448	0,449	0,491	0,498	0,477	0,528	0,508	0,500	0,460	0,480
B5	0,689	0,509	0,464	0,507	1	0,478	0,459	0,453	0,456	0,510	0,486	0,515	0,509	0,475	0,480
B6	0,660	0,495	0,500	0,448	0,478	1	0,484	0,489	0,473	0,461	0,437	0,504	0,441	0,474	0,459
B7	0,661	0,495	0,466	0,449	0,459	0,484	1	0,490	0,457	0,461	0,475	0,474	0,446	0,450	0,467
B8	0,672	0,461	0,441	0,491	0,453	0,489	0,490	1	0,442	0,504	0,486	0,424	0,435	0,479	0,469
B9	0,661	0,434	0,428	0,498	0,456	0,473	0,457	0,442	1	0,421	0,454	0,475	0,477	0,492	0,436
B10	0,683	0,521	0,442	0,477	0,510	0,461	0,461	0,504	0,421	1	0,502	0,449	0,481	0,472	0,440
B11	0,690	0,465	0,437	0,528	0,486	0,437	0,475	0,486	0,454	0,502	1	0,521	0,450	0,505	0,483
B12	0,689	0,481	0,499	0,508	0,515	0,504	0,474	0,424	0,475	0,449	0,521	1	0,463	0,484	0,494
B13	0,699	0,496	0,522	0,500	0,509	0,441	0,446	0,435	0,477	0,481	0,450	0,463	1	0,484	0,478
B14	0,678	0,512	0,459	0,460	0,475	0,474	0,450	0,479	0,492	0,472	0,505	0,484	0,484	1	0,483
B15	0,651	0,481	0,433	0,480	0,480	0,459	0,467	0,469	0,436	0,440	0,483	0,494	0,478	0,483	1

* Все коэффициенты значимы на уровне больше 0,01.

Факторный анализ

Для матрицы корреляций, представленной в табл. 2, использовался факторный анализ с помощью метода главных компонент¹, результаты которого приводятся в табл. 3.

Как видно из таблицы 3, четыре первых фактора объясняют почти 67% вариации. Такой процент объясненной дисперсии, как правило, считается вполне приемлемым при использовании факторного анализа в социологических и маркетинговых исследованиях. Табл. 3 показывает также, что данный факторный анализ не является особенно удачным, поскольку общности демонстрируют неравномерность в объяснении дисперсии отдельных переменных (особенно для переменных В14, В15 по сравнению с переменной В1), и, по всей видимости, было бы целесообразно увеличить число факторов. Однако, учитывая «модельность» примера, мы не будем этого делать, тем более что, по нашим наблюдениям, на значения общностей исследователи внимания чаще всего вообще не обращают.

На рис. 1 представлены гистограммы распределения построенных индивидуальных значений факторов. Как видно из рис. 1, общий вид распределений напоминает нормальные кривые.

Обратим внимание еще на одну распространенную ошибку в интерпретации результатов факторного анализа. Как правило, исследователи устанавливают определенную «точку отсечения» для значений факторных нагрузок и значения меньшие этой точки в интерпретации не участвуют. Рассмотрим, к примеру, матрицу, приведенную в табл. 3. Если установить в качестве такой «точки отсечения» значение 0,3, то для объяснения, скажем, 3-го фактора будут использоваться переменные В9, В10, В13. Далее индивидуальные значения 3-го фактора будут рассматриваться именно как индекс, характеризующий поведение данных трех переменных.

¹ Строго говоря, метод главных компонент не является методом факторного анализа, однако мы использовали его сознательно, поскольку именно этот метод чаще всего применяется при решении прикладных задач.

Таблица 3

МАТРИЦА ФАКТОРНЫХ НАГРУЗОК, ПРОЦЕНТ ОБЪЯСНЕННОЙ ДИСПЕРСИИ, ОБЩНОСТИ ДЛЯ МОДЕЛЬНЫХ ДАННЫХ

	Факторы				Общности
	1	2	3	4	
V1	0,959	-0,007	-0,016	-0,020	0,920
V2	0,730	0,054	0,219	-0,218	0,631
V3	0,699	0,457	0,116	-0,068	0,716
V4	0,728	-0,171	-0,301	-0,036	0,652
V5	0,729	0,020	-0,092	-0,260	0,609
V6	0,711	0,224	0,260	0,277	0,699
V7	0,704	-0,007	0,280	0,245	0,634
V8	0,705	-0,318	0,293	0,164	0,710
V9	0,692	0,071	-0,296	0,340	0,687
V10	0,714	-0,249	0,232	-0,326	0,731
V11	0,723	-0,334	-0,177	0,020	0,665
V12	0,728	0,145	-0,220	0,117	0,614
V13	0,719	0,243	-0,143	-0,312	0,694
V14	0,721	-0,039	-0,042	0,059	0,527
V15	0,704	-0,080	-0,092	0,051	0,513
Процент объясненной дисперсии	53,8%	4,4%	4,3%	4,2%	

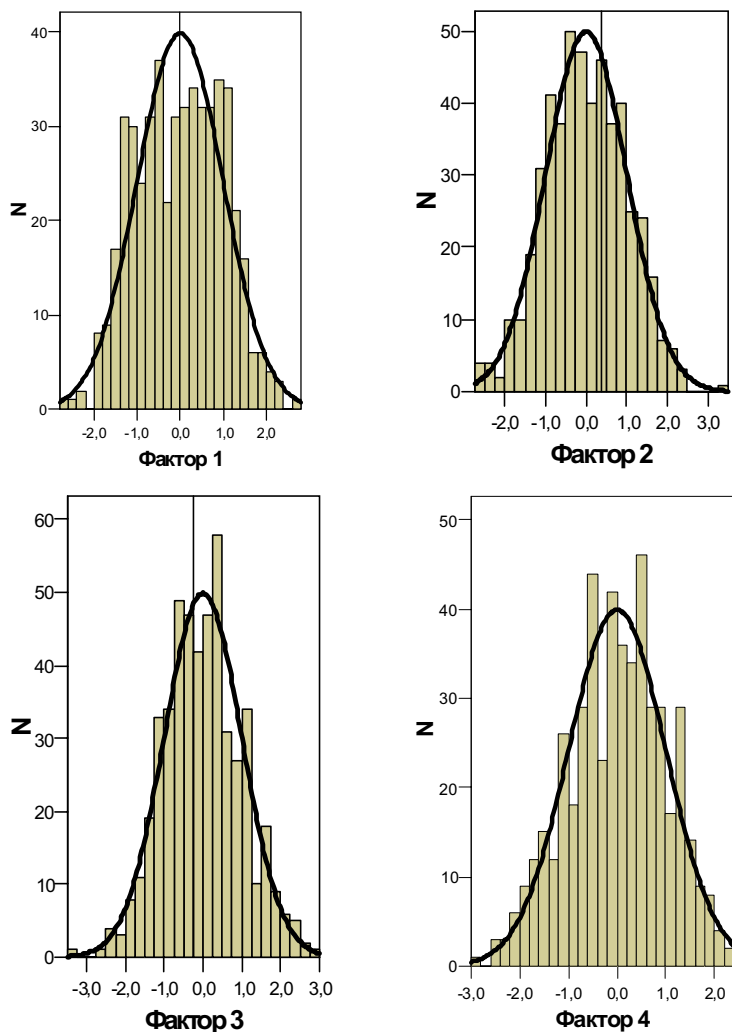


Рис. 1. Гистограммы распределения индивидуальных значений первых 4 факторов для анализа главных компонент табл. 3.

Однако при вычислении индивидуальных значений фактора используются не только те переменные, которые легли в основу интерпретации, но и все остальные (хотя и, разумеется, с меньшими весами). Проблема состоит в том, что хотя веса остальных переменных и меньше, однако самих этих переменных гораздо больше и, соответственно, их суммарный вклад в полученные значения фактора достаточно велик. Таким образом, построенный фактор, который интерпретируется на основе включенных в рассмотрение переменных, становится индексом, отражающим поведение совсем иных переменных.

Для иллюстрации этой мысли оценим то, на сколько значения 3-го фактора, которые вычисляет SPSS при использовании стандартной процедуры сохранения факторов, определяются теми переменными, которые участвовали в интерпретации данного фактора (В9, В10, В13). Для решения этой задачи построим регрессионную модель, в которой в качестве зависимой переменной выступает 3-й фактор, а в качестве независимых переменных – переменные, которые легли в основу интерпретации фактора (В9, В10, В13).

Коэффициент R^2 , определяющий качество такой модели, в нашем примере составил 0,26. Иными словами, лишь 26% поведения 3-го фактора объясняются теми тремя переменными, которые используются для интерпретации этого фактора.

Таблица регрессионных коэффициентов (табл. 4) демонстрирует еще один любопытный факт. В матрице факторных нагрузок (табл. 3) переменные В9, В10, В13 для 3-го фактора имеют достаточно близкие по абсолютной величине значения нагрузок и, следовательно, при объяснении поведения этого фактора будет предполагаться, что три рассматриваемые переменные имеют на данный фактор приблизительно одинаковое влияние. Однако значения регрессионных коэффициентов показывают, что переменная В13 воздействует на построенный фактор гораздо слабее, чем переменные В9 и В10.

Таким образом, традиционная интерпретация поведения 3-го фактора как индекса, отражающего поведение переменных В9, В10 и В13, дает абсолютно неадекватную картину. Во-первых, эти три переменные объясняют лишь 26% поведения фактора. Во-вторых, степень влияния данных переменных на построенный фактор не может быть объяснена на основе матрицы факторных нагрузок.

Таблица 4

РЕГРЕССИОННЫЕ КОЭФФИЦИЕНТЫ МОДЕЛИ ДЛЯ ОЦЕНКИ
ВЛИЯНИЯ ПЕРЕМЕННЫХ В9, В10, В13 НА 3-Й ФАКТОР

	Нестандарти- зованные коэффициенты		Стандарти- зованные коэффи- циенты	<i>t</i>	Значимость
	<i>B</i>	Ст. ошибка	Beta		
Константа	0,055	0,047		1,178	0,239
В9	-0,015	0,002	-0,418	-9,235	0,000
В10	0,019	0,002	0,495	10,911	0,000
В13	-0,007	0,002	-0,182	-3,882	0,000

Иерархический кластерный анализ

В качестве первого метода классификации созданного модельного массива используем иерархический метод кластерного анализа с разбиением на два кластера. Переменными будут выступать индивидуальные значения 4 построенных факторов. В качестве параметров кластеризации выбираются те, которые предлагаются SPSS по умолчанию.

Дендрограмма, построенная программой иерархического кластерного анализа, не позволяет увидеть две группы, которые заданы в модельном массиве. При разбиении массива на 2 кластера результат получается абсолютно неудовлетворительный – в одном

кластере оказывается 1 объект, а во втором кластере – 499. Даже разбиение на 7 кластеров показывает, что массив разделяется на один большой кластер, два средних и четыре мелких. При этом принадлежность объектов, исходно принадлежащих двум заданным группам по построенным кластерам, достаточно произвольна (табл. 5).

Таблица 5

КОЛИЧЕСТВО ОБЪЕКТОВ ИЗ 2 МОДЕЛЬНЫХ ГРУПП,
РАЗНЕСЕННОЕ ПО 10 КЛАСТЕРАМ
(КЛАСТЕРИЗАЦИЯ НА 4 ФАКТОРАХ)

Номера кластеров	Исходные группы		Всего
	1	2	
1	237	211	448
2	6	18	24
3	2	20	22
4	0	1	1
5	3	0	3
6	1	0	1
7	1	0	1
Всего	250	250	500

Табл. 5 показывает, что применение иерархического кластерного анализа для выделения двух модельных групп не дает хоть сколько-нибудь приемлемого результата.

Попробуем провести иерархический кластерный анализ, используя в качестве переменных не построенные ранее факторы, а непосредственно 15 исходных переменных. Результат такой кластеризации представлен в табл. 6. Как показывает таблица, полученную классификацию можно вполне признать удовлетворительной, поскольку лишь 60 (около 12%) объектов были отнесены к неверным группам.

Таблица 6

КОЛИЧЕСТВО ОБЪЕКТОВ ИЗ 2 МОДЕЛЬНЫХ ГРУПП,
РАЗНЕСЕННОЕ ПО 2 КЛАСТЕРАМ
(КЛАСТЕРИЗАЦИЯ НА ИСХОДНЫХ ПЕРЕМЕННЫХ)

Номера кластеров	Исходные группы		Всего
	1	2	
1	205	15	220
2	45	235	280
Всего	250	250	500

Кластерный анализ методом K-средних

Представленная в SPSS команда «K-means» (K-средних) является гораздо более технологичной по сравнению с программой иерархического кластерного анализа, и, соответственно, используется гораздо чаще. Вначале проведем разбиение модельного массива на 2 кластера на построенных ранее факторах, не задавая начальные центры кластеров. Соответствие исходных групп построенным объектам представлено в табл. 7.

Таблица 7

КОЛИЧЕСТВО ОБЪЕКТОВ ИЗ 2 МОДЕЛЬНЫХ ГРУПП,
РАЗНЕСЕННОЕ ПО 2 КЛАСТЕРАМ
(КЛАСТЕРИЗАЦИЯ НА 4 ФАКТОРАХ)

Номера кластеров	Исходные группы		Всего
	1	2	
1	174	99	273
2	76	151	227
Всего	250	250	500

Результат кластеризации трудно признать удовлетворительным, поскольку почти 35% объектов были классифицированы ошибочно.

Кластеризация с помощью алгоритма K -средних при использовании в качестве переменных не построенных факторов, а непосредственно исходных показателей, дает гораздо более приемлемые результаты (табл. 8). При таком разбиении менее 9% объектов классифицируются ошибочно.

Таблица 8

КОЛИЧЕСТВО ОБЪЕКТОВ ИЗ 2 МОДЕЛЬНЫХ ГРУПП,
РАЗНЕСЕННОЕ ПО 2 КЛАСТЕРАМ
(КЛАСТЕРИЗАЦИЯ НА ИСХОДНЫХ ПЕРЕМЕННЫХ)

Номера кластеров	Исходные группы		Всего
	1	2	
1	220	13	233
2	30	237	267
Всего	250	250	500

Обсуждение результатов

Может создаться впечатление, что основной причиной недопустимо низкого качества кластеризации наших модельных данных при использовании индивидуальных значений факторов как переменных является плохая исходная факторная модель. Действительно, представленная в табл. 3 матрица факторных нагрузок весьма неудобна для интерпретации. В результате мы имеем факторы, как показано на примере 3-го фактора, с невысокими факторными нагрузками, т.е. слабо связаны с исходными переменными. Когда же выяснилось, что три переменные, выбранные для интерпретации 3-го фактора, объясняют его лишь на 26%, было трудно ожидать хороших результатов от кластеризации на факторах.

Традиционно, для улучшения (скорее – упрощения) матрицы факторных нагрузок используют вращение факторной матрицы. В табл. 9 приведена матрица факторных нагрузок после вращения матрицы табл. 3 методом Варимакс.

Таблица 9

МАТРИЦА ФАКТОРНЫХ НАГРУЗОК ПОСЛЕ ВРАЩЕНИЯ
ВАРИМАКС, ПРОЦЕНТ ОБЪЯСНЕННОЙ ДИСПЕРСИИ, ОБЩНОСТИ
ДЛЯ МОДЕЛЬНЫХ ДАННЫХ

	Факторы				Общности
	1	2	3	4	
В9	0,704				0,687
В4	0,660				0,652
В11	0,618				0,665
В12	0,583				0,614
В1	0,555				0,920
В15	0,503				0,513
В14					0,527
В13		0,697			0,694
В3		0,674			0,716
В5		0,523			0,609
В2		0,503			0,631
В10			0,729		0,731
В8			0,591	0,515	0,710
В6				0,707	0,699
В7				0,637	0,634
Процент объясненной дисперсии	20,5%	16,1%	15,3%	14,8%	

* В матрице не приводятся факторные нагрузки меньше 0,5.

После проведенного вращения ситуация несколько улучшилась. Коэффициент R^2 показывает, что 3-й фактор объясняется переменными В8 и В10 почти на 60%. Однако остаются отмеченные ранее недостатки интерпретации поведения фактора как индекса, отражающего выделенные переменные, основанной на матрице факторных нагрузок. Так, регрессионный коэффициент

при переменной В8 в два раза меньше коэффициента при переменной В10, хотя факторные нагрузки у этих переменных отличаются лишь на 20%.

Не спасает вращение матрицы факторных нагрузок и при решении задачи кластеризации объектов, основанной на значениях факторов. Так применение алгоритма *K*-средних к факторам, полученным по результатам ортогонального вращения, дает такое же решение, как и разбиение на кластеры, основанное на факторном анализе без вращения (табл. 7).

Другим возможным объяснением плохого качества кластеризации на факторах может быть то, что факторная модель (неважно, с вращением или без) объясняет далеко не всю дисперсию исходных признаков (в рассматривавшемся примере – 67%). Соответственно, построенные факторы включают лишь 2/3 исходной информации переменных, и, следовательно, кластеризация получается низкого качества из-за потери значительной части исходной информации. Однако это объяснение является несостоятельным.

Мы повторили эксперимент с модельным массивом данных, выделив не 4, как ранее, а 10 факторов. Очевидно, что такой факторный анализ становится гораздо хуже интерпретируемым, но зато он объясняет более 88% дисперсии исходных переменных. Кажется, что качество кластеризации, основанной на значениях таких 10 факторов, должно быть близким к кластеризации на исходных переменных (табл. 8), и, по крайней мере, должно быть лучше, чем качество кластеризации, основанной на 4 факторах (табл. 7). На самом деле качество кластеризации на 10 факторах посредством применения метода *K*-средних гораздо хуже, чем качество кластеризации на 4 факторах. Количество ошибочно классифицированных объектов при использовании индивидуальных значений 10 факторов составляет 56%, притом, что для случая 4 факторов этот показатель был равен 35%.

Причиной выявленных «странностей» является то, что все предлагаемые в традиционных статистических пакетах (SPSS,

STATISTICA и др.) методы факторного анализа строят ортогональные факторы¹. Далее в случае использования в факторном анализе нескольких десятков переменных полученные индивидуальные значения факторов имеют, как правило, распределения достаточно близкие к нормальному (за исключением случаев тех факторов, которые имеют очень высокие нагрузки для небольшого числа переменных). Таким образом, если взглянуть на полученный массив переменных (факторов), которые подвергаются кластеризации, то мы увидим, что это данные из независимых переменных с многомерным нормальным распределением. Ясно, что кластеризация такого массива все равно может быть проведена, поскольку нет таких данных, которые нельзя кластеризовать. Другое дело, что полученный результат будет иметь вполне случайный характер, и его качество будет определяться лишь интерпретационными способностями исследователя. Вообще «замечательность» таких эвристических методов, как факторный и кластерный анализы, состоит в том, что качество получаемых с их помощью результатов верифицируется лишь критерием «правдоподобности», что целиком находится в руках исследователя.

¹ Мы не рассматриваем здесь сюжеты неортогонального вращения факторов.