Препринт статьи для летнего номера "Социология 4М"

Привет Асхат Хасянович!

Написал тут (грешным делом) статью на ту тему, о которой мы с тобой неоднократно говорили. Статья пойдет в летний номер "Социология 4М". Мне было бы интересно твое мнение и, возможно, советы по поводу неточностей и глупостей.

С уважением,
А.Крыштановский

Крыштановский А.О.[1]

«Кластеры на факторах» - об одном распространенном заблуждении

Задача построения классификации единиц исследования является весьма распространенной как в социологических, так и в маркетинговых исследованиях. Получение ***однородных групп*** объектов (чаще всего – респондентов), то есть таких групп, которые приблизительно одинаково ведут себя в одинаковых ситуациях (чаще всего – 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, кластерный анализ методом К-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 «непонятным» 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.

[1] Крыштановский А.О. Заведующий кафедрой методов сбора и анализа социологической информации Государственного университета – Высшая школа экономики, декан факультета социологии ГУ-ВШЭ.

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 привести к сильному смещению результирующей кластеризации, а классифицируем объекты по 3-4 переменным (факторам), каждая из которых при этом имеет более или менее вразумительную интерпретацию.

Данный подход, по моим наблюдениям, достаточно широко используется в практике как социологических, так и маркетинговых исследований. Такой путь рекомендуется и в достаточно широко распространенной книге А. Бююля и Н.Цефеля[2]. К сожалению, 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-х групп респондентов на 15 вопросов. Файл синтакисиса 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-х 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-х группах достаточно сильно различаются между собой (Таблица 1), и, следовательно, можно считать, что эти 2

[2] А. Бююль, П.Цефель. SPSS: искусство обработки информации. DiaSoft, М-Ст-Петербург-Киев, 2002, с. 394-398. В данной главе факторный анализ назван почему-то «факториальным», но это, по всей видимости, пробел редактуры.
Отметим, что такой же подход пропагандируется авторами и в разделе, в котором обсуждается кластерный анализ методом К-средних (с. 404-409).

группы представляют собой существенно различные совокупности объектов, что, по логике исследования, должно обнаружиться с помощью методов классификации.

Таблица 1

Статистические характеристики модельных переменных в 2-х группах значения

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

* Средние значение всех переменных в 2-х группах различаются с вероятностью P>0,99

Таблица 2 демонстрирует значения коэффициентов корреляции для созданных 15-ти переменных.

Таблица 2

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	,688	,662	,700	,689	,660	,661	,672	,661	,683	,690	,689	,699	,678	,651
B2	,688	1	,479	,478	,509	,495	,495	,461	,434	,521	,465	,481	,496	,512	,481
B3	,662	,479	1	,452	,464	,500	,466	,441	,428	,442	,437	,499	,522	,459	,433
B4	,700	,478	,452	1	,507	,448	,449	,491	,498	,477	,528	,508	,500	,460	,480
B5	,689	,509	,464	,507	1	,478	,459	,453	,456	,510	,486	,515	,509	,475	,480

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

* Все коэффициенты значимы на уровне P> 0.01.

Факторный анализ

Для матрицы корреляций, представленной в таблице 2 проведен факторный анализ с помощью метода главных компонент[3], результаты которого приводятся в Табл. 3.

Таблица 3

Матрица факторных нагрузок, процент объясненной дисперсии, общности для модельных данных

	Факторы				Общности
	1	2	3	4	
B1	,959	-,007	-,016	-,020	,920
B2	,730	,054	,219	-,218	,631
B3	,699	,457	,116	-,068	,716
B4	,728	-,171	-,301	-,036	,652
B5	,729	,020	-,092	-,260	,609
B6	,711	,224	,260	,277	,699
B7	,704	-,007	,280	,245	,634
B8	,705	-,318	,293	,164	,710
B9	,692	,071	-,296	,340	,687
B10	,714	-,249	,232	-,326	,731
B11	,723	-,334	-,177	,020	,665
B12	,728	,145	-,220	,117	,614
B13	,719	,243	-,143	-,312	,694
B14	,721	-,039	-,042	,059	,527
B15	,704	-,080	-,092	,051	,513
Процент объясненной дисперсии	**53,8%**	**4,4%**	**4,3%**	**4,2%**	

Как видно из таблицы 3, четыре первых фактора объясняют почти 67% информации. Такой процент объясненной дисперсии, как правило, считается вполне приемлемым при использовании факторного анализа в социологических и

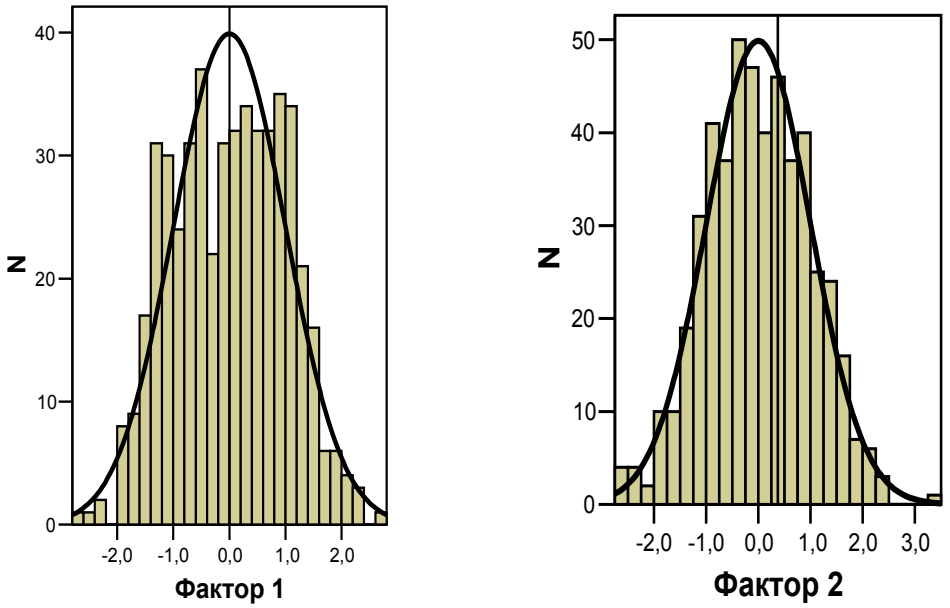
[3] Строго говоря, метод главных компонент не является методом факторного анализа, однако мы использовали его сознательно, поскольку именно этот метод чаще всего используется при решении прикладных задач.

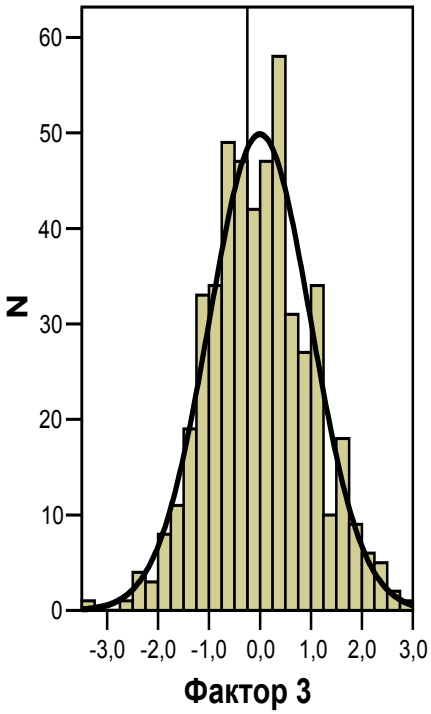
маркетинговых исследованиях. Таблица 3 показывает также, что данный факторный анализ не является особенно удачным, поскольку общности демонстрируют неравномерность в объяснении дисперсии отдельных переменных (особенно для переменных B14, B15, по сравнению с переменной B1), и, по всей видимости, было 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.

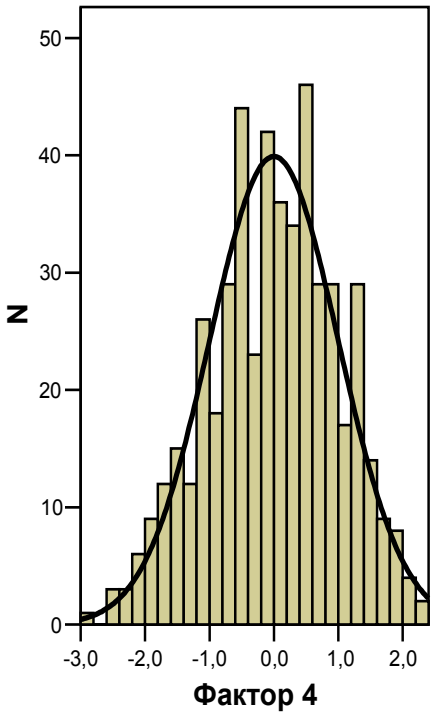
На рисунке 1 представлены гистограммы распределения построенных индивидуальных значений факторов. Как видно из рисунка 1, общий вид распределений напоминает нормальные кривые.

Рисунок 1

Гистограммы распределения индивидуальных значений первых 4-х факторов, для анализа главных компонент Таблицы 3







Обратим внимание еще на одну распространенную ошибку в интерпретации результатов факторного анализа. Как правило, при интерпретации исследователи устанавливают определенную «точку отсечения» для значений факторных нагрузок и значения меньшие этой точки в интерпретации не участвуют. Рассмотрим, к примеру, матрицу, приведенную в таблице 3. Если установить в качестве такой «точки отсечения» значение 0,3, то для интерпретации, скажем, 3-го фактора будут использоваться переменные B9, B10, B13. Далее, индивидуальные значения 3-го фактора будут рассматриваться именно как индекс, характеризующий поведение данных трех переменных.

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 (B9, B10, B13). Для решения этой задачи построим регрессионную

модель, в которой в качестве зависимой переменной выступает 3-й фактор, а в качестве независимых переменных – переменные, легшие в основу интерпретации фактора (B9, B10, B13).

Коэффициент $\mathbf{R^2}$, 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 (таблица 4) демонстрирует еще один любопытный факт. В матрице факторных нагрузок (Табл. 3) переменные B9, B10, B13 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 B13 влияет на построенный фактор гораздо слабее, чем переменные B9 и B10.

Таким образом, традиционная интерпретация поведения 3-го фактора как индекса, отражающего поведение переменных B9, B10 и B13, дает абсолютно неадекватную картину. Во-первых, эти три переменных 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 B9, B10, B13 на 3-й фактор

	Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Значимость
	B	Ст. ошибка	Beta		
Константа	,055	,047		1,178	,239
B9	-,015	,002	-,418	-9,235	,000
B10	,019	,002	,495	10,911	,000
B13	-,007	,002	-,182	-3,882	,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-х 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 (таблица 5).

Таблица 5

Количество объектов из 2-х модельных групп,
разнесенное по 10-ти кластерам
(кластеризация на 4-х 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

Таблица 5 показывает, что использование иерархического кластерного анализа для выделения двух модельных групп не дает хоть сколь-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 таблице 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-х модельных групп,
разнесенное по 2-м 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 К-средних.

Представленная в SPSS команда «K-means» (К-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 таблице 7.

Таблица 7.

Количество объектов из 2-х модельных групп,
разнесенное по 2-м кластерам
(кластеризация на 4-х 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 К-средних при использовании в качестве переменных не построенных факторов, а непосредственно исходных показателей, дает гораздо более приемлемые результаты (таблица 8). При таком разбиении менее 9% объектов классифицируются ошибочно.

Таблица 8.

Количество объектов из 2-х модельных групп,
разнесенное по 2-м 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 в качестве

переменных индивидуальных значений факторов является плохая исходная факторная модель. Действительно, представленная в Табл. 3 матрица факторных нагрузок весьма неудобна для интерпретации. В результате мы имеем факторы, которые, как показано на примере 3-го фактора, с невысокими факторными нагрузками, то есть 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 таблице 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	
B9	,704				,687
B4	,660				,652
B11	,618				,665
B12	,583				,614
B1	,555				,920
B15	,503				,513
B14					,527
B13		,697			,694
B3		,674			,716
B5		,523			,609
B2		,503			,631
B10			,729		,731
B8			,591	,515	,710
B6				,707	,699
B7				,637	,634
Процент объясненной дисперсии	**20,5%**	**16,1%**	**15,3%**	**14,8%**	

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

После проведенного вращения ситуация несколько улучшилась. Коэффициент $\mathbf{R^2}$ показывает, что 3-й фактор объясняется переменными B8 и B10 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 B8 в два раза меньше коэффициента при переменной B10 (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 К-средних к факторам, полученным по результатам ортогонального вращения, дает точно 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, уж, по крайней мере, должно быть лучше, чем качество кластеризации на, основанной на 4-х факторах (Табл.7). На самом деле качество кластеризации на 10 факторах при использовании метода К-средних гораздо хуже, чем качество кластеризации на 4-х факторах. Количество ошибочно классифицированных объектов при использовании индивидуальных значений 10 факторов составляет 56%, при том, что для случая 4-х 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 факторы[4]. 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.

[4] Мы не рассматриваем здесь сюжеты неортогонального вращения факторов.

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.