

«Кластеры на факторах»: pro et contra

Юдин Г.Б., MA in Sociology

(Государственный Университет – Высшая Школа Экономики)

При обработке социологических данных может возникать задача совмещения факторного и кластерного анализа. Эта задача возникает обычно в тех случаях, когда сначала исходная информация преобразуется путём выделения факторов, а затем на основе преобразованных данных производится поиск в изучаемой совокупности отдельных друг от друга классов. Существует несколько подходов к совмещению двух видов анализа.

Наиболее распространённый подход заключается в проведении факторного анализа на исходных переменных, оценке значений факторных шкал для отдельных респондентов и использовании этих значений для построения классификации. Такая двух-этапная схема получила название «тандемного подхода» (tandem approach) и может применяться исходя из различных формальных и содержательных соображений. Однако за ней стоит комбинация технических процедур, составляющих кластерный и факторный анализ. Таким образом, результаты совмещения методов в некоторой степени предопределяются последствиями этого сочетания, которые нередко скрыты от взгляда исследователя.

На опасность применения кластерного анализа по результатам факторного в отечественной литературе обратил внимание А. Крыштановский [1]. Он провёл эксперимент на смоделированных данных и заключил, что даже в случае исходно хорошего разделения данных на группы факторный анализ сглаживает различия. Кластерный анализ на основе значений факторов для респондентов даёт значительную долю ошибок классификации, не распознавая чёткую структуру, заложенную в исходные данные.

Объяснение выявленным проблемам Крыштановский видит, во-первых, в том, что стандартный факторный анализ даёт независимые факторы, а во-вторых – в том, что значения факторов имеют распределение, похожее на нормальное. Соответственно, на вход в кластерный анализ поступают данные по ортогональным признакам с многомерным нормальным распределением. Такие данные не содержат разбиения на классы, и получившееся на выходе разбиение совершенно произвольно. Результаты, которые

дают на таких данных методы иерархического анализа и особенно анализа к-средних, являются чистыми артефактами.

В зарубежной литературе наиболее жёсткую позицию по отношению к тандемному подходу заняли Ф. Араби и Л. Хьюберт, назвавшие его «устаревшей и статистически безосновательной» практикой [3, p.167]. Другие авторы занимают более умеренную позицию, признавая, что при определённых обстоятельствах применение этого подхода может быть целесообразным [4;5;7]. В целом, можно выделить следующие преимущества тандемного метода: а) уменьшение размерности пространства; б) стандартизация шкал; в) выравнивание веса шкал. Третье преимущество связано с тем, что если в исходных данных могут быть выделены факторы и эти факторы нагружены неодинаковым числом переменных, то без предварительного факторного анализа одни переменные будут в ходе кластеризации иметь больший вес, чем другие [2].

Важный недостаток сочетания подходов состоит в том, что в результате факторного анализа часть информации по малозначимым переменным выпадает из интерпретации, в то время как кластеризация включает эту информацию; в итоге два метода дают неконсистентные результаты. Второе возражение апеллирует к тому, что факторный анализ уменьшает расстояния между объектами, имеющиеся в исходных данных, и тем самым снижает устойчивость кластеризации.

Существуют альтернативные варианты совмещения двух методов. Среди них следует выделить репрезентацию факторов исходными переменными – в этом случае для кластерного анализа используются не значения факторных шкал, но значения двух-трёх переменных, сильнее всего связанных с каждым из выделенных факторов [5]. Также существует возможность взвешивания исходных переменных – по результатам факторного анализа переменным, измеряющим один и тот же латентный признак, присваиваются коэффициенты, понижающие их значимость при классификации. Или, напротив, переменные, указывающие на уникальные признаки, могут быть повторены в массиве с целью увеличения их веса [4]. Наконец, не следует забывать о возможности проведения кластерного анализа на исходных переменных, без выделения факторов. Все эти техники дают результаты, отличные от тандемного подхода.

В дискуссии о правомерности применения тандемного подхода ключевым становится вопрос о том, какова степень негативного воздействия, которое факторный

анализ оказывает на исходные данные для кластеризации. Иными словами, задача состоит в том, чтобы выяснить, в какой мере распределение данных приближается за счёт факторного анализа к многомерному нормальному. Две процедуры, осуществляемые факторным анализом - ортогонализация и нормализация, оцениваются с помощью экспериментов.

В случаях, когда такие эксперименты проводятся на специально сгенерированных данных [1;4], экспериментальная процедура вводит некоторые предположения, которые плохо согласуются с реальными социологическими данными. Использование нормального распределения для создания модельных данных чревато существенными искажениями. Как показывает обсуждение проблемы идентификации нормального распределения, актуальной для такого направления анализа данных, как моделирование смесей (mixture modeling), наложение нормальных распределений даёт распределение, неотличимое от нормального [6]. Таким образом, представляется рискованным осуществлять моделирование двухкластерной структуры через совмещение двух нормальных распределений с различными параметрами. Выход можно найти в том, чтобы при создании модельных данных разделять два нормальных распределения очень большим интервалом (например, в пять стандартных отклонений [6, p. 365]). Однако в исключительно редких ситуациях с такой выраженной неоднородностью использование кластерного анализа вообще оказывается избыточным.

Нормальность распределения, полученного совмещением двух нормальных распределений, иллюстрируется Рисунком 1 и подтверждается тестом Колмогорова-Смирнова¹. Следует напомнить, что для кластерного анализа это ключевой вопрос, поскольку построение таксономии напрямую зависит от формы распределения.

¹ Мы воспроизвели данные Крыштановского по предложенному им алгоритму. Уровень значимости для теста Колмогорова-Смирнова $\alpha > 0,9$.

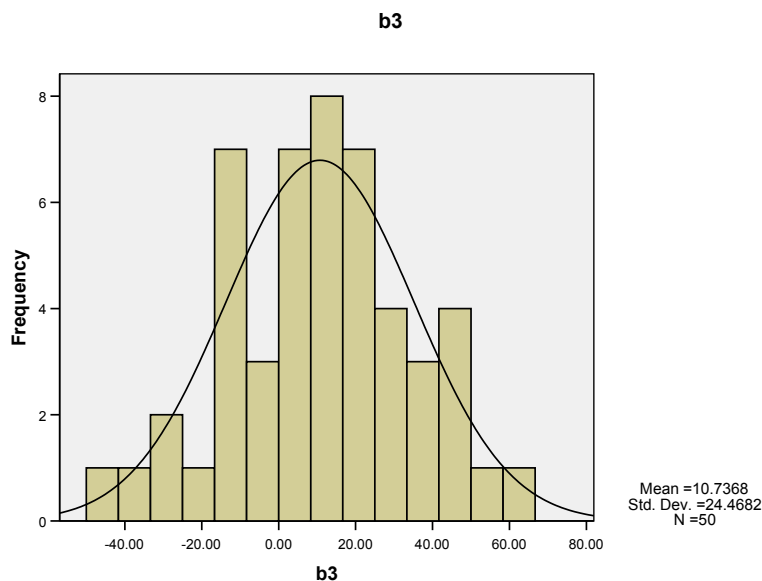


Рисунок 1. Пример распределения смеси двух нормальных распределений (модельные данные).

Если же отказаться от такого рода моделирования и оценить негативное воздействие факторного анализа на реальных данных, оно окажется куда менее заметным. Поскольку переменные в социологических исследованиях не так часто имеют нормальное распределение, получаемые на их основе факторы также не подчиняются нормальному закону. Мы провели такого рода проверку на данных World Values Survey по России (1999 год, N=2500). В качестве тестируемых переменных были использованы показатели доверия граждан к различным институтам (четырёхбалльные шкалы, всего – 15 институтов)². Для 15 институтов был проведён предварительный эксплораторный факторный анализ. Затем был проведён конфирматорный факторный анализ с 5 факторами; переменные распределялись по факторам в соответствии с результатами эксплораторного анализа и исходя из содержательных соображений. 5 факторов выглядят следующим образом:

Фактор 1: Доверие системе здравоохранения, системе правосудия, крупным компаниям;

Фактор 2: Доверие ЕС, НАТО, ООН;

Фактор 3: Доверие профсоюзам, парламенту, прессе;

² В ходе факторного анализа использовалось вращение varimax.

Фактор 4: Доверие вооружённым силам, системе образования, милиции;

Фактор 5: Доверие церкви, государственным учреждениям (службам), системе социального обеспечения.

Более тщательное решение содержательных задач потребовало бы реструктуризации факторов, однако для наших целей это не имеет определяющего значения. Качество факторного решения удовлетворительное: факторы объясняют не менее 56% дисперсии вошедших в них признаков. Конструирование факторных шкал производилось по методу регрессии³. Фактор выделялся двумя способами: отбором одной главной компоненты и методом максимального правдоподобия. Значимых различий между двумя этими способами не зафиксировано.

На Рисунке 2 мы приводим распределения двух из трёх исходных переменных, которые легли в основу первого фактора. Согласно тесту Колмогорова-Смирнова, эти распределения значимо отличаются от нормального⁴.

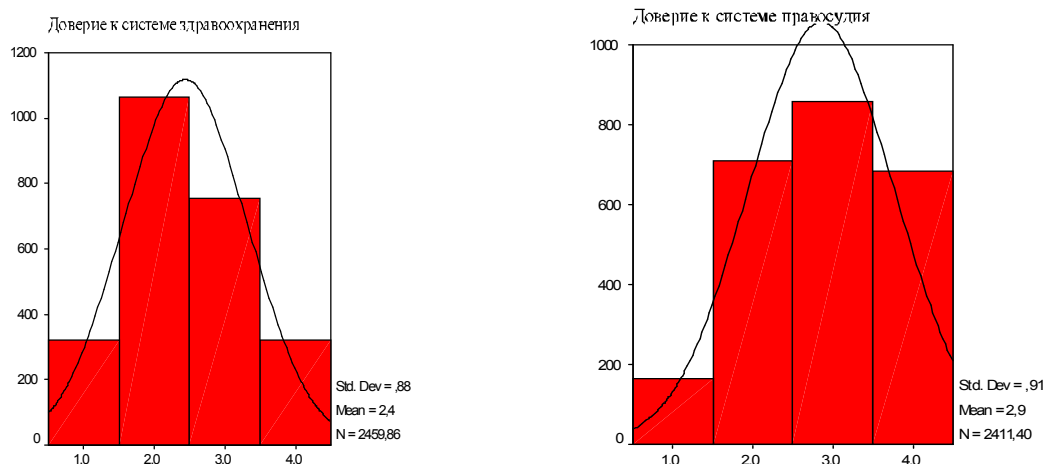


Рисунок 2. Распределение двух из пятнадцати исходных переменных (данные WVS-99).

На Рисунке 3 приведены распределения двух из пяти факторов, выделенных

³ Заметим, что этот метод, как правило, даёт неортогональные факторы даже при обычном эксплораторном анализе методом главных компонент. В данном случае, поскольку применялся конфирматорный анализ, корреляции между факторами оказались значительными (на уровне 0,30-0,55).

⁴ Уровень значимости $\alpha=0,000$ во всех случаях. В значительной степени отклонение от нормальности обусловлено порядковой шкалой, на которой измерены переменные. Однако мы не обсуждаем здесь проблемы факторного анализа на переменных, имеющих распределение, отличное от нормального. В подавляющем числе случаев социологу приходится иметь дело с ненормально распределёнными данными. Также мы воздерживаемся от обсуждения правомерности проведения факторного анализа на порядковых шкалах. Такие методы, как латентно-классовый анализ и оптимальное шкалирование являются более адекватными с точки зрения теории измерений. Тем не менее, вопрос о том, в какой мере эти методы могут заменить рассматриваемые здесь методы, заслуживает отдельного обсуждения.

методом максимального правдоподобия. Тест Колмогорова-Смирнова также не выявил сходства с нормальным распределением ни в одном из пяти случаев⁵.

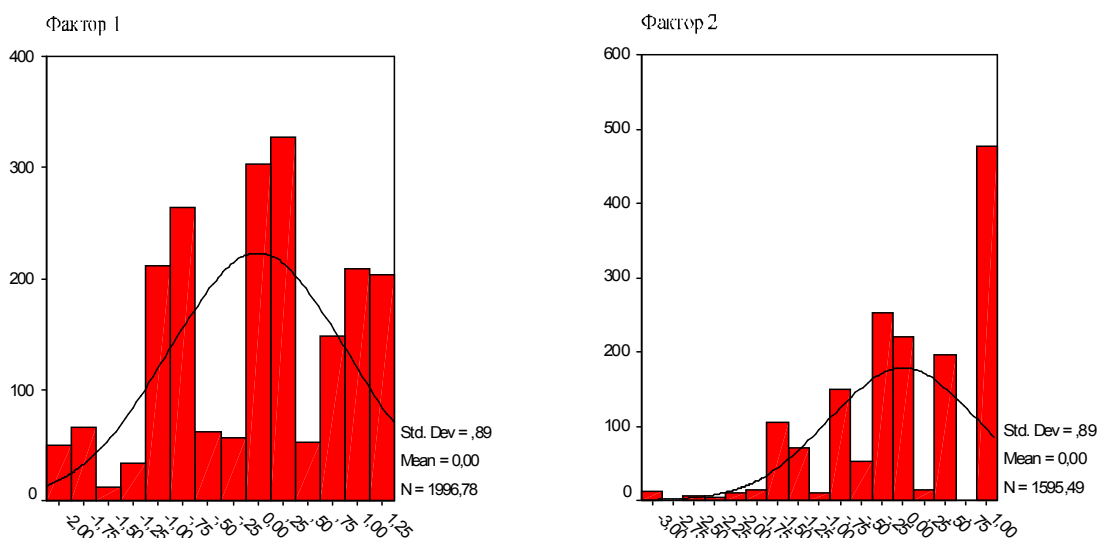


Рисунок 3. Распределение двух из пяти полученных факторов (данные WVS-99).

В этой ситуации ортогональность факторов не имеет определяющего значения. Однако если данные близки к нормальным, вероятно, имеет смысл прибегать к косугольным факторным решением и конфирматорному факторному анализу, который позволяет получать более полную интерпретацию матрицы факторных нагрузок⁶.

Таким образом, возражения против тандемного подхода страдают чрезмерным обобщением. По-видимому, вывод о применимости тандемного подхода целесообразно делать исходя из особенностей имеющихся данных [4]; главным образом, из их распределения. Противники «кластеров на факторах» правы, указывая на опасность, которую факторный анализ таит для кластеризации. Преобразования, производимые факторным анализом, действительно оказывают негативное воздействие на процедуру кластеризации, однако в реальной практике это воздействие может быть незначительным. Эксперименты, проводимые на реальных данных, показывают, что применение тандемного

⁵ Уровень значимости $\alpha=0,000$ во всех случаях. Можно отметить, что рассматриваемые переменные центрированы: в некоторых случаях это несколько облегчает дальнейшую работу с полученными факторами.

⁶ Впрочем, здесь следует отметить, что использование конфирматорного анализа несколько нивелирует преимущества тандемного подхода, поскольку допускает существование серьезных корреляций между факторами. Если в исходном наборе переменных есть большой блок переменных, сильно связанных между собой, эти переменные вновь получают большой вес при кластеризации. Решение о использовании конфирматорного анализа должно приниматься с учётом корреляционной структуры исходных переменных. Автор признателен Николаю Бабичу за это замечание.

подхода может быть разумным [5]. Этот подход имеет преимущества не только перед кластеризацией исходных данных, но и перед другими техниками (взвешиванием переменных; заменой факторов на переменные с наибольшими нагрузками).

В заключение следует заметить, что слепое использование факторного анализа как средства стандартизации или снижения размерности пространства чревато неадекватными выводами. Отсутствие содержательной модели для латентных переменных повышает риск деструктивного воздействия факторного анализа на кластеризацию. Напротив, наличие такой модели может не только сделать применение тандемного подхода более оправданным, но и существенно обогатить интерпретацию.

Литература.

1. Крыштановский А.О. «Кластеры на факторах» - об одном распространённом заблуждении // Социология: 4М. 2005. № 21. с. 172-187.
2. Aldenderfer M., Blashfield R. Cluster analysis. Beverly Hills, CA, 1984.
3. Arabie P., Hubert L. Cluster analysis in marketing research // Advanced methods of marketing research. Oxford, 1994.
4. Elder J. Knowing when to factor: Simulating the tandem approach to cluster analysis // Proceeding of the Sawtooth Software Conference. 1999. p. 101-108.
5. Fiedler J., McDonald J. Market figmentation: Clustering on factor scores versus individual variables / Paper presented to the AMA Advanced Research Techniques Forum. 1993.
6. Rindskopf D. Mixture or homogenous? // Psychological methods. 2003. №3. p. 364-368.
7. Schaffer C., Green P. Cluster-based market segmentation: Some further comparisons of alternative approaches // Journal of the Market Research Society. 1998. №2. p. 155-163.