УДК 316:004.9 DOI 10.18698/2306-8477-2017-10-473

Факторный анализ и категориальный метод главных компонент: сравнительный анализ и практическое применение для обработки результатов анкетирования

© Е.Е. Фомина

Тверской государственный технический университет, Тверь, 170026, Россия

Анкетирование представляет собой один из основных инструментов изучения состояния общественного мнения в работе социолога. Первичным результатом анкетирования является, как правило, база данных, требующая последующего глубокого анализа и поиска взаимосвязей между исследуемыми показателями. Для решения этой задачи могут быть применены факторный анализ и категориальный метод главных компонент, которые позволяют придать содержательный смысл полученным результатам. Несмотря на то что с применением данных методов решают одну задачу, в них используются различные алгоритмы выделения интегральных характеристик, поэтому проблема выбора подходящего метода является актуальной. В статье проведено сравнение факторного анализа и категориального метода главных компонент как с теоретической позиции, так и с точки зрения практического применения. Рассмотрен пример обработки результатов анкетирования, предложены методические рекомендации.

Ключевые слова: *факторный анализ, алгоритм CatPCA, метод главных компонент, анкетирование*

Анкетирование — метод исследования, позволяющий оперативно осуществлять мониторинг состояния и тенденций изменения общественного мнения по тем или иным вопросам.

Основное достоинство метода заключается в том, что исследователь может опросить большое число респондентов, проживающих в разных регионах, и получить сопоставимые данные, для анализа которых удобно использовать методы математической статистики [1].

Обработка анкет — многоэтапная процедура, включающая как выполнение рутинных, механических операций, так и решение содержательных задач, получение обоснованных выводов.

На практике для обработки анкет наиболее часто используются следующие методы:

- расчет показателей описательной статистики;
- подсчет распределения ответов в зависимости от значений дополнительных переменных, таких как пол, возраст, образование и др.;
- построение таблиц сопряженности и проверка статистических гипотез о независимости признаков с использованием критерия хи-квадрат;
- выявление корреляционной зависимости между отдельными признаками;

- графическая обработка информации.

Наряду с вышеуказанными, важную роль также играют методы и алгоритмы интерпретации, позволяющие придать содержательный смысл результатам анкетирования. К ним относятся *факторный анализ* (ФА) и *категориальный метод главных компонент* (CatPCA — Categorical Principal Component Analysis), которые направлены на решение следующих задач:

- поиск скрытых закономерностей во множестве исследуемых переменных, которые возникают вследствие воздействия на них некоторых факторов;
- изучение статистической взаимосвязи между признаками и выделенными факторами;
- описание предметной области с помощью общих факторов, количество которых намного меньше, чем исходное число переменных [2, 3].

Несмотря на то, что методы направлены на решение одинаковых задач, каждый из них имеет свои особенности реализации, поэтому встает актуальная проблема выбора. Цель настоящей статьи — провести сравнение ФА и CatPCA, а также рассмотреть особенности их реализации для решения задачи обработки данных, полученных в результате анкетирования.

Факторный анализ. ФА — класс процедур многомерного статистического анализа, направленный на выявление латентных переменных (факторов), отвечающих за наличие линейных статистических связей (корреляций) между наблюдаемыми переменными [4].

ФА основывается на предположении, что исследуемое явление, определяемое некоторой системой признаков, изменяющихся согласованно, может быть описано с помощью меньшего числа других латентных переменных, называемых *факторами,* объясняющими причины этих изменений. Число факторов намного меньше числа исходных переменных.

Факторы — это группы определенных переменных, коррелирующих между собой больше, чем с переменными, входящими в другой фактор. Таким образом, содержательный смысл факторов может быть выявлен путем исследования корреляционной матрицы исходных данных.

Например, при изучении ценностных предпочтений какой-либо социальной группы необходимо установить наличие взаимосвязей среди большого числа параметров (пола, возраста, образования, различных групп ценностных ориентаций и т. д.). Для исследования всех возможных зависимостей между этими переменными потребовалось бы рассчитать и проанализировать большой набор коэффициентов корреляций. Вместо этого можно заменить исходный набор

признаков меньшим числом латентных переменных или факторов, не поддающихся непосредственному измерению (например, факторами карьерного роста, социальной активности, духовности и морально-этических ценностей и др.). Предполагается, что выделенные факторы являются наиболее существенными и определяющими.

Математическая модель факторного анализа представляет собой набор линейных уравнений, в котором каждая наблюдаемая переменная x_i выражается в виде линейной комбинации общих факторов $F_1, F_2, ..., F_n$ и уникального фактора U_i:

$$x_i = \sum_{k=1}^{n} a_{ik} F_k + U_i ,$$

где x_i — переменная, $i = \overrightarrow{1, m}$, ($m$ — количество переменных); n — количество факторов $(n \ll m)$; a_{ik} — факторная нагрузка; F_k — общий фактор, $k = \overrightarrow{1, n}$; U_i — частный фактор.

Процедура ФА включает в себя три этапа.

Этап 1. Построение корреляционной матрицы системы переменных путем расчета коэффициентов линейной корреляции Пирсона. Причем корреляционная матрица может быть представлена не в исходном, а в редуцированном виде, т. е. на ее главной диагонали будут стоять не единицы, а оценки общих нормированных дисперсий, рассчитываемые по методу наибольшей корреляции или по методу Барта [5]. Использование редуцированной матрицы объясняется тем, что в ФА дисперсия признаков может быть объяснена не на 100 %, а несколько меньше с учетом существования частных факторов.

Этап 2. Извлечение факторов и расчет факторных нагрузок a_{ik}, являющихся основным предметом интерпретации. На этом этапе используют методы компонентного анализа (метод главных компонент), главных факторов и максимального правдоподобия.

На практике для выделения факторов наиболее часто используется *метод главных компонент* (МГК). Его основная идея заключается в том, чтобы выделить в многомерном пространстве $X = (x_1, x_2, ..., x_k)$ группы тесно коррелирующих между собой переменных и заменить их без потери информативности главными компонентами $Y = (y_1, y_2, ..., y_m)$. Математическая модель МГК может быть записана в виде

$$y_j = \sum_{i=1}^{k} \alpha_{ij} z_i ,$$

где y_j — главная компонента ($j = \overrightarrow{1, m}$); α_{ij} — коэффициент, отражающий вклад переменной z_i в главную компоненту y_i; z_i — стандартизированная исходная переменная $z_i = (x_i - \overline{x_i}) / s_i$, $i = \overrightarrow{1, k}$.

Выделение главных компонент осуществляется по представленному ниже алгоритму.

1. Стандартизация исходных переменных, приводящая к тому, что дисперсии всех стандартизированных переменных становятся одинаковыми (все стандартизированные переменные имеют одинаковую информативность) и начало координат переносится в центр облака данных.

2. Линейное преобразование пространства $Z = (z_1, z_2, ..., z_k)$ с целью построения нового ортогонального пространства главных компонент $Y = (y_1, y_2, ..., y_k)$:

$$y_j = \sum_{i=1}^{k} \alpha_{ij} z_i, \quad i, j = \overrightarrow{1,k}.$$

Для осуществления этого преобразования необходимо рассчитать коэффициенты $\Lambda = \{\alpha_{ij}\}$. Они определяются исходя из следующих требований:

- главные компоненты должны быть линейными комбинациями переменных $z_1, z_2, ..., z_k$;
- главные компоненты должны быть ортогональными;
- первая главная компонента должна иметь максимальную выборочную дисперсию, вторая главная компонента должна иметь максимальную выборочную дисперсию при фиксированной первой и т.д.: $s^2(y_1) \geq s^2(y_2) \geq ... \geq s^2(y_k)$;
- суммарная дисперсия исходных переменных должна быть равна суммарной дисперсии главных компонент.

Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений $(\lambda_1, \lambda_2, ..., \lambda_k)$ корреляционной матрицы исходных данных. При этом собственные числа будут равны дисперсиям новых переменных $s^2(y_1) = \lambda_1 \geq s^2(y_2) = \lambda_2 \geq ... \geq s^2(y_k) = \lambda_k$, а собственные векторы будут совпадать со столбцами матрицы $\Lambda = \{\alpha_{ij}\}$: $(\alpha_{1i} ... \alpha_{ki})^T$ — i-й собственный вектор, соответствующий собственному числу λ_i.

Значения α_{ij} называются факторными нагрузками. Они представляют собой коэффициенты корреляции между исходными переменными и главными компонентами. Факторы включают в себя те переменные, для которых $|\alpha_{ij}| > 0,7$.

3. Сокращение размерности пространства $Y = (y_1, y_2, ..., y_k)$ посредством отсечения неинформативных переменных. Для решения этой задачи используются:

- критерий Кайзера, связанный с собственными значениями: в число главных компонент включают только те переменные, которым соответствуют собственные значения $\lambda_i \geq 1$, так как их информативная ценность выше;
- критерий, связанный с долей сохраненной дисперсии: суммарная дисперсия главных компонент должна быть не менее заданной доли;
- критерий Кеттела (критерий «каменистой осыпи»), согласно которому собственные числа отображаются на графике, где по оси абсцисс откладываются их номера, а по оси ординат — значения. Далее ищется точка на графике, где убывание собственных чисел максимально замедляется. Номер, соответствующий этому числу, и определяет оптимальное количество факторов.

Этап 3. Вращение факторного решения, которое используется в том случае, если выделенные факторы невозможно достаточно наглядно интерпретировать [2, 3, 6]. На практике используют следующие методы вращения: варимакс, квартимакс, эквимакс, биквартимакс.

Ограничения метода. ФА осуществляется по коррелированным переменным. Его основным объектом исследования является корреляционная матрица, построенная с использованием коэффициента линейной корреляции Пирсона. Следовательно, анализируемые данные должны подчиняться многомерному нормальному закону распределения; значения признаков необходимо измерить в интервальной шкале. Допускается также анализ порядковых переменных с большим числом значений, так как порядковые шкалы с высоким числом градаций обладают свойствами интервальных шкал [7]. Однако в анкетах эти требования часто не выполняются: анализируемые переменные имеют разный уровень измерений (в порядковых, номинальных и дихотомических шкалах). Применение ФА для таких переменных может привести к искажению факторной структуры, которое связано с искажением коэффициента корреляции. В этом случае альтернативой ФА выступает CatPCA.

Метод CatPCA. Данный метод предназначен для решения задачи снижения размерности пространства исходных данных, измеренных в любых шкалах.

Он обладает большими возможностями, в частности, при обработке результатов анкетирования, так как не накладывает никаких ограничений на тип переменных. CatPCA позволяет одновременно анализировать как количественные переменные, измеренные в интервальных, порядковых, номинальных, дихотомических шкалах, так и качественные переменные. Кроме того, с помощью данного метода можно решить проблему пропущенных данных, так как отсутствующая информация по какой-либо из переменных воспринимается как

самостоятельная категория или как отдельное для каждого объекта значение.

Обработка данных методом CatPCA включает в себя два этапа. На первом этапе происходит процедура оцифровки переменных, которая опирается на принципы оптимального шкалирования; на втором этапе выполняется редукция размерности данных.

Математическая формализация метода имеет следующий вид. Рассмотрим матрицу исходных переменных $X = (x_1, x_2, ..., x_p)$ размерности $n \times p$, где переменная $x_j \in R^n$ и может принимать L_j различных значений. Для нее требуется определить матрицу интегральных характеристик Z таким образом, чтобы функция $\sigma(Z,W)$ принимала минимальное значение [8]:

$$\sigma(Z,W) = \sum_{j=1}^{p} tr(Z - G_j W_j)^T (Z - G_j W_j) \to \min \qquad (1)$$

при ограничениях

$$Z^T 1_n = 0_r; \qquad (2)$$

$$Z^T Z = nI_r, \qquad (3)$$

т. е. интегральные характеристики должны удовлетворять условиям центрированности и ортонормированности.

Здесь Z — матрица интегральных характеристик размерности $n \times r$; G_j — матрица индикаторов размерности $n \times L_j$ для исходной переменной x_j;

$$G_j(i, l_j) = \begin{cases} 1, \text{если объект } i \text{ относится к категории } l_j, \\ 0 \text{ в противном случае;} \end{cases}$$

W_j — матрица размерности $L_j \times p$ переменной x_j, содержащая координаты всех ее категорий в r-м пространстве; 1_n — единичный вектор размерности $n \times 1$; 0_r — единичный вектор размерности $r \times 1$; I_r — единичная матрица размером $r \times r$.

Оптимизация функции (1) при ограничениях (2), (3) осуществляется с помощью итерационного алгоритма *Princals* [8].

Оцифровка переменных в алгоритме CatPCA происходит таким образом, что собственные значения компонент, рассчитанные по матрице корреляций оптимизированных переменных, максимизируются.

Показателем, позволяющим оценить качество проведенного анализа, является альфа Кронбаха (α) — коэффициент, показывающий внутреннюю согласованность характеристик, описывающих один объект. Альфа Кронбаха лежит в интервале от $-\infty$ до 1. Если $\alpha \succ 0{,}7$,

то модель считается качественной. Когда в модели появляются факторы с собственным значением меньше 1, α становится отрицательным. Следовательно, оптимальным является то число факторов, при котором α принимает положительное значение [9].

Сравнительный анализ методов ФА и CatPCA представлен в табл. 1.

Таблица 1

Сравнительный анализ методов ФА и CatPCA

Показатель	Факторный анализ	CatPCA
1. Ограничения метода		
1.1. Нормальность распределения исходных данных	Предполагается	Не требуется
1.2. Уровень измерения анализируемых данных	Интервальная шкала или порядковая с большим числом градаций	Любой
1.3. Использование переменных, измеренных на разных уровнях	Не предполагается	Возможно
2. Исходные данные для анализа		
2.1. Предварительное преобразование данных	Не предполагается	Оцифровка
2.2. Исходные данные для выделения факторов	Матрица корреляций исходных переменных	Матрица корреляций «оцифрованных» переменных
3. Критерий отбора оптимального числа факторов	Критерий Кайзера; критерий, связанный с долей сохраненной дисперсии; критерий Кеттела (критерий «каменистой осыпи»)	Альфа Кронбаха
4. Возможность вращения	Предполагается	Не предполагается

Пример. Рассмотрим пример практического применения ФА и CatPCA. Методы использовались для обработки результатов анкетирования, целью которого было установить степень доверия жителей России к органам власти, степень удовлетворенности экономической, политической ситуацией и отношение к мигрантам.

Анкета и база данных с результатами анкетирования были взяты с сайта http://sophist.hse.ru/ (единый архив экономических и социологических данных). Объем выборки составил 1000 человек. В анализируемых данных частота каждой категории по всем вопросам включает больше восьми наблюдений, что обеспечивает устойчивость применяемых методов.

Автоматизированная обработка полученных данных осуществлялась в пакете SPSS.

Анкета

А1. Насколько Вы интересуетесь политикой?

Варианты ответа: 1 — «очень интересуюсь»; 2 — «интересуюсь в некоторой степени»; 3 — «мало интересуюсь»; 4 — «совсем не интересуюсь»; 5 — «затрудняюсь ответить».

А2. Насколько Вы доверяете Парламенту нашей страны?

Варианты ответа: от 0 — «совершенно не доверяю» до 10 — «полностью доверяю».

А3. Насколько Вы доверяете судебно-правовой системе нашей страны?

Варианты ответа: от 0 — «совершенно не доверяю» до 10 — «полностью доверяю».

А4. Насколько Вы доверяете полиции нашей страны?

Варианты ответа: от 0 — «совершенно не доверяю» до 10 — «полностью доверяю».

А5. Насколько Вы доверяете политикам в нашей стране?

Варианты ответа: от 0 — «совершенно не доверяю» до 10 — «полностью доверяю».

А6. Насколько Вы доверяете политическим партиям в нашей стране?

Варианты ответа: от 0 — «совершенно не доверяю» до 10 — «полностью доверяю».

А7. Насколько Вы доверяете Европейскому парламенту?

Варианты ответа: от 0 — «совершенно не доверяю» до 10 — «полностью доверяю».

А8. Насколько Вы доверяете ООН?

Варианты ответа: от 0 — «совершенно не доверяю» до 10 — «полностью доверяю».

В1. Насколько Вы удовлетворены своей жизнью в целом?

Варианты ответа: от 0 — «совершенно не удовлетворен» до 10 — «полностью удовлетворен».

В2. Насколько Вы удовлетворены нынешним состоянием экономики?

Варианты ответа: от 0 — «совершенно не удовлетворен» до 10 — «полностью удовлетворен».

В3. Насколько Вы удовлетворены тем, как руководство страны выполняет свою работу?

Варианты ответа: от 0 — «совершенно не удовлетворен» до 10 — «полностью удовлетворен».

В4. Насколько Вы удовлетворены тем, как работает демократия в нашей стране?

Варианты ответа: от 0 — «совершенно не удовлетворен» до 10 — «полностью удовлетворен».

В5. Как Вы оцениваете нынешнее состояние системы образования в нашей стране?

Варианты ответа: от 0 — «очень плохое» до 10 — «очень хорошее».

В6. Как Вы оцениваете нынешнее состояние системы здравоохранения в нашей стране?

Варианты ответа: от 0 — «очень плохое» до 10 — «очень хорошее».

В7. Правительство должно принять меры для уменьшения разницы в доходах между людьми?

Варианты ответа: от 0 — «полностью согласен» до 10 — «совершенно не согласен».

М1. Следует ли позволить людям той же национальности, что и большинство населения страны, переезжать жить в нашу страну?

Варианты ответа: 1 — «следует позволить многим таким людям переезжать»; 2 — «можно позволить некоторым таким людям переезжать»; 3 — «следует позволить переезжать лишь немногим из них»; 4 — «никому не разрешать».

М2. А людям, которые по национальности или расовой принадлежности отличаются от большинства населения страны?

Варианты ответа: 1 — «следует позволить многим таким людям переезжать»; 2 — «можно позволить некоторым таким людям переезжать»; 3 — «следует позволить переезжать лишь немногим из них»; 4 — «никому не разрешать».

М3. А если говорить о людях из более бедных стран за пределами Европы?

Варианты ответа: 1 — «следует позволить многим таким людям переезжать»; 2 — «можно позволить некоторым таким людям переезжать»; 3 — «следует позволить переезжать лишь немногим из них»; 4 — «никому не разрешать».

М4. То, что люди из других стран переезжают в нашу страну, в целом хорошо или плохо сказывается на экономике страны?

Варианты ответа: от 0 — «плохо для экономики» до 10 — «хорошо для экономики».

М5. Приток людей из других стран скорее разрушает или обогащает культуру нашей страны?

Варианты ответа: от 0 — «разрушает культуру нашей страны» до 10 — «обогащает культуру нашей страны».

М6. С притоком людей из других стран наша страна как место для жизни становится лучше или хуже?

Варианты ответа: от 0 — «становится хуже» до 10 — «становится лучше».

Результаты применения ФА. Для обработки данных применялся ФА, где в качестве метода выделения факторов использовался МГК.

На первом этапе решался вопрос об оптимальном количестве факторов. Для этого был использован критерий Кайзера. В итоге выделено шесть факторов (шесть собственных чисел больше единицы [2, 3]), оказывающих влияние на результаты анкетирования (см. табл. 1).

Как можно видеть из табл. 2, общий процент дисперсии, объясняемый всеми факторами, равен 59,1.

Таблица 2

Собственные значения ФА

Значение	Фактор					
	1	2	3	4	5	6
Собственное значение	4,90	2,10	1,80	1,30	1,10	1,10
Процент общей дисперсии	23,50	10,30	9,10	6,40	5,00	4,80
Кумулятивный процент	23,50	33,80	42,90	49,30	54,30	59,10

На следующем этапе был проведен расчет и анализ факторных нагрузок. Нужно отметить, что первоначальная матрица факторных нагрузок не позволила выделить четкую факторную структуру, поэтому для более наглядной интерпретации решения был применен метод вращения *Варимакс исходных*. Корреляция между фактором и переменной считалась сильной, если модуль факторной нагрузки принимал значение больше 0,70. Результаты расчетов представлены в табл. 3.

Таблица 3

Факторные нагрузки, полученные с использованием метода главных компонент и метода вращения Варимакс исходных

Переменная	Факторы					
	1	2	3	4	5	6
A2	0,79	—	—	—	—	—
A3	0,70	—	—	—	—	—
A4	0,78	—	—	—	—	—
A5	0,81	—	—	—	—	—
A6	0,78	—	—	—	—	—
M1	—	0,82	—	—	—	—
M2	—	0,88	—	—	—	—
M3	—	0,81	—	—	—	—
M4	—	—	0,71	—	—	—
M5	—	—	0,71	—	—	—
M6	—	—	0,76	—	—	—
A7	—	—	—	0,85	—	—
A8	—	—	—	0,85	—	—
B3	—	—	—	—	0,71	—
A1	—	—	—	—	—	0,71

Анализ данных табл. 2 показывает, что первый, т. е. генеральный, фактор включает в себя пять переменных с положительными значениями коэффициента корреляции: A2...A6. При ответе на эти вопросы респонденты выражали свое доверие к политической, правовой и судебной системе в России. Условно фактор можно обозначить как «доверие». В целом факторный вес составил 23,5 %.

Второй фактор тесно связан с вопросами M1...M3, в которых респонденты отмечали свое отношение к мигрантам, приезжающим в РФ. Условно фактор можно обозначить как «миграция». Его факторный вес составил 10,3 %.

Третий фактор включает в себя вопросы M4...M6, в которых респонденты оценивали влияние мигрантов на экономику и культуру России. Условно фактор можно обозначить как «влияние мигрантов». Его факторный вес составил 9,1 %.

Четвертый фактор связан с вопросами A7...A8, в которых респонденты высказывали степень доверия Европейскому парламенту и ООН. Условно фактор можно обозначить как «доверие к ООН и Европарламенту». Его факторный вес составил 6,40 %.

Пятый и шестой факторы плохо поддаются объяснению, так как включают в себя только по одному признаку.

Таким образом, построена модель, которая является вполне адекватной с точки зрения ее интерпретации.

Устойчивость полученного решения проверялась путем применения аналогичной процедуры к контрольным выборкам. Результаты

применения ФА к исследуемой выборке и контрольным подвыборкам практически не отличаются. Процент объясненной дисперсии варьируется от 57,7 до 59,1, т. е. выделенные факторы являются устойчивыми.

Результаты применения CatPCA. При обработке данных методом CatPCA был определен порядковый уровень оптимального шкалирования и пропущенные значения заменены модой.

Первым шагом применения метода являлось определение оптимального числа факторов. Для этого была построена таблица с результатами вычисления значения альфы Кронбаха (табл. 4) для числа размерностей, равного 10. Анализ данных в табл. 4 показывает, что выделяются девять компонент с собственными значениями больше 1. Учитывая, что начиная с размерности, равной 5, альфа Кронбаха становится достаточной малой величиной (меньшей 0,3), в методе CatPCA была выбрана конечная размерность, равная 4, и процедура расчета проведена заново. Сводка по модели представлена в табл. 5. Построенная модель объясняет 66,3 % общей дисперсии, что на 17,0 % больше, чем в ФА (для четырехфакторной модели). Общее значение коэффициента альфа Кронбаха равно 0,97, следовательно, построенная модель хорошо описывает исходные данные.

Таблица 4

Сводка для модели с десятью компонентами

Значение	Размерность										
	1	2	3	4	5	6	7	8	9	10	Всего
Альфа Кронбаха	0,88	0,7	0,53	0,51	0,28	0,26	0,22	0,12	0,09	– 0,15	0,99
Собственное значение	6,43	3,04	2,02	1,94	1,36	1,32	1,26	1,13	1,09	0,87	20,51

Таблица 5

Сводка для модели с четырьмя компонентами

Значение	Размерность				
	1	2	3	4	Всего
Альфа Кронбаха	0,89	0,70	0,53	0,50	0,97
Собственное значение	6,90	3,03	2,04	1,93	13,91
Объясненная дисперсия, %	32,90	14,50	9,70	9,20	66,30

Анализ нагрузок (см. табл. 5) приводит к следующей интерпретации полученных результатов. Первая компонента включает в себя пять переменных: A2...A6, т. е. она собрала в себе вопросы, касающиеся доверия к политической, правовой и судебной системам. Этот результат полностью соответствует результату ФА и совпадает с первым фактором.

Вторая компонента связана с переменными М1...М3, отражающими отношение к мигрантам. Она также полностью совпадает со вторым фактором.

Третья и четвертая компоненты имеют невысокие нагрузки (в табл. 6 курсивом отмечены нагрузки, больше 0,46). Отсутствие процедуры вращения построенного решения не дает возможности попытаться устранить этот недостаток. Третья компонента связана с вопросами, касающимися удовлетворенностью жизнью в целом, состоянием экономики и демократии. Нужно отметить, что аналогичная компонента отсутствует в факторной структуре, построенной методом ФА. Четвертая компонента совпадает с третьим фактором, полученным в процедуре ФА, и отражает мнение респондентов о влиянии мигрантов на экономику и культуру страны.

Таблица 6

Нагрузки для метода CatPCA

Переменная	Компонента			
	1	2	3	4
А2	0,77	—	—	—
А3	0,78	—	—	—
А4	0,77	—	—	—
А5	0,82	—	—	—
А6	0,79	—	—	—
М1	—	0,71	—	—
М2	—	0,82	—	—
М3	—	0,79	—	—
В1	—	—	*0,46*	—
В2	—	—	*0,49*	—
В4	—	—	*0,49*	—
М4	—	—	—	*0,55*
М5	—	—	—	*0,60*
М6	—	—	—	*0,64*

Устойчивость построенного решения проверялась путем применения метода к контрольным подвыборкам. Результаты выделения компонент в контрольных и исследуемой подвыборках аналогичные.

Применение рассмотренных выше методов для решения практической задачи анализа результатов анкетирования показало, что они строят похожие факторные модели, однако процент объясненной дисперсии в модели, полученной в результате применения CatPCA, больше, чем в ФА. Это происходит из-за разных подходов к обработке пропущенных данных, дополнительной оцифровки переменных в алгоритме CatPCA, а также уровня измерения исходных данных.

Выводы. Целью проведенного исследования было сопоставление процедур ФА и метода CatPCA, которые сравнивались как с теоретической позиции, так и с точки зрения практического применения при обработке анкет.

Хотя ФА и CatPCA решают одну и ту же задачу редукции размерности пространства исходных переменных, они работают с разными исходными данными, используют различные алгоритмы выделения факторов и различные критерии для определения их оптимального числа.

ФА работает с матрицей корреляций исходных переменных, применяет для выделения факторов набор специальных алгоритмов и методов. Важным преимуществом ФА является возможность вращения факторной структуры, которое позволяет получить более наглядную интерпретацию имеющегося решения. Существенным ограничением метода является уровень измерения анализируемых данных.

Возможности метода CatPCA гораздо шире, так как он не накладывает никаких ограничений на исходные данные. Однако в данном алгоритме не предусмотрено вращение построенного решения, что в ряде случаев может затруднить интерпретацию результата.

Таким образом, при решении практических задач в первую очередь необходимо обращать внимание на тип переменных. Если данные имеют различный уровень измерений, а также преобладают порядковые, дихотомические или ранговые переменные с небольшим числом градаций, то основным инструментом анализа будет являться метод CatPCA. Применение ФА в таких случаях недопустимо, так как он может привести к искажению факторной структуры, связанному с неверным расчетом коэффициента корреляции.

При анализе интервальных или ранговых переменных с большим числом (более 5) допустимо использование как ФА, так и CatPCA. В большинстве случаев методы будут давать близкое решение. При этом необходимо обращать внимание на оптимальное число интегральных характеристик, структуру полученного решения, возможность его интерпретации (большую роль на этом этапе играет процедура вращения решения, предусмотренная в ФА), процент объясненной дисперсии. Сравнительный анализ полученных результатов позволит отдать предпочтение тому или иному методу.

ЛИТЕРАТУРА

[1] Беседа, интервьюирование и анкетирование. *Характеристика методов исследования.* URL: http://cito-web.yspu.org/link1/metod/met93/node4.html (дата обращения 01.09.2016).

[2] Бессокирная Г.П. Факторный анализ: традиции использования и новые возможности. *Социология: методология, методы, математическое моделирование*, 2000, № 12, с. 142–153.

[3] Буреева Н.Н. Многомерный статистический анализ с использованием ППП «STATISTICA». Нижний Новгород, НГУ им. Н.И. Лобачевского, 2007, 112 с.
[4] Общая психология. В кн.: Петровский А.В., ред. *Энциклопедический словарь. В 6 т. Т. 1*. Москва, ПЕР СЭ, 2005, 251 с.
[5] *Общий алгоритм и теоретические проблемы факторного анализа*. URL: http://www.studfiles.ru/preview/1938850/page:3/ (дата обращения 01.03.2017).
[6] Фомина Е.Е. Применение факторного анализа для обработки результатов анкетирования. *Социосфера*, 2016, № 3, с. 122–127.
[7] Толстова Ю.Н. *Измерение в социологии*. Москва, КДУ, 2007, 288 с.
[8] *CATPCA Algorithms*. URL: https://www.ibm.com/support/knowledgecenter/zh/SSLVMB_21.0.0/com.ibm.spss.statistics.help/alg_catpca.htm (дата обращения 03.04.2017).
[9] *Сравнительный анализ методов категориального факторного анализа*. URL: https://www.hse.ru/data/2013/06/07/1283920122/%D0%92%D0%9A%D0%A0%20%D0%A8%D0%98%D0%A8%D0%9A%D0%9E.doc (дата обращения 20.03.2017).

Статья поступила в редакцию 07.06.2017

Ссылку на эту статью просим оформлять следующим образом:
Фомина Е.Е. Факторный анализ и категориальный метод главных компонент: сравнительный анализ и практическое применение для обработки результатов анкетирования. *Гуманитарный вестник*, 2017, вып. 10.
http://dx.doi.org/10.18698/2306-8477-2017-10-473

Фомина Елена Евгеньевна — канд. техн. наук, доцент кафедры «Информатика и прикладная математика» Тверского государственного технического университета. Область научных интересов — математическое моделирование в медицине и социологии. e-mail: f-elena2008@yandex.ru

Factor analysis and categorial principal component analysis: comparative analysis and practical application for processing of questionnaire survey results

© E.E. Fomina

Tver State Technical University, Tver, 170026, Russia

The questionnaire survey is one of the main tools of studying the state of public opinion in the work of the sociologist. The primary result of the survey is usually a database that requires further in-depth analysis and search for relationships among variables being studied. To solve this problem a factor analysis and categorial principal component analysis can be applied. It allows making obtained results meaningful. Despite the fact that using these methods one problem is solved, they include different algorithms for determining integral characteristics, so the problem of selecting the appropriate method is relevant. The article compares factor analysis and categorial principal component analysis both from a theoretical position and from the point of view of practical application. An example of processing of questionnaire survey results is considered, and methodological recommendations are offered.

Keywords: *factor analysis, CatPCA algorithm, principal component analysis, questionnaire survey*

REFERENCES

[1] Shcherbak A.P. Beseda, intervyuirovanie i anketirovanie [Conversation, interviewing and questioning] In: *Kharakteristika metodov issledovaniya* [Characteristics of research methods]. Rybinsk, Rybinskiy filial Yaroslavskogo Gosudarstvennogo Pedagogicheskogo Universiteta im. K.D. Ushinskogo Publ., 2007. Available at: http://cito-web.yspu.org/link1/metod/met93/node4.html (accessed September 01, 2016).

[2] Bessokirnaya G.P. *Sotsiologiya: metodologiya, metods, matematicheskoe modelirovanie — Sociology: Methodology, Methods, Mathematical Modeling*, 2000, no. 12, pp. 142–153.

[3] Bureeva N.N. *Mnogomernyy statisticheskiy analiz s ispolzovaniem paketa prikladnykh program "STATISTICA"* [Multivariate statistical analysis using the "STATISTICA" software package]. Nizhniy Novgorod, Lobachevsky Nizhegorodskiy Gosudarstvennyy Universitet Publ., 2007, 112 p.

[4] Obshchaya psikhologiya [General Psychology]. In: Petrovsky A.V., red. *Entsiklopedicheskiy slovar. V 6 tomakh. Tom 1* [Petrovsky A.V., ed. Encyclopedic dictionary. In 6 vols. Vol. 1]. Moscow, PER SE Publ., 2005, 251 p.

[5] *Obshchiy algoritm i teoreticheskie problemy faktornogo analiza* [General algorithm and theoretical problems of the factor analysis]. Available at: http://www.studfiles.ru/preview/1938850/page:3/ (accessed March 01, 2017).

[6] Fomina E.E. *Sotsiosfera — Sociosphere*, 2016, no. 3, pp. 122–127.

[7] Tolstova Yu.N. *Izmereniya v sotsioligii* [Measurement in Sociology]. Moscow, KDU Publ., 2007, 288 p.

[8] *CatPCA Algorithms*. Available at: https://www.ibm.com/support/knowledgecenter/zh/SSLVMB_21.0.0/com.ibm.spss.statistics.help/ alg_catpca.htm (accessed April 03, 2017).

[9] *Sravnitelnyy analiz metodov kategorialnogo faktornogo analiza* [Comparative analysis of methods of categorical factor analysis]. Available at: https://www.hse.ru/data/2013/06/07/1283920122/%D0%92%D0%9A%D0%A0%20%D0%A8%D0%98%D0%A8%D0%9A%D0%9E.doc (accessed March 20, 2017).

Fomina E.E., Cand. Sc. (Eng.), Associate Professor, Department of Informatics and Applied Mathematics, Tver State University. Research interests: mathematical modeling in medicine and sociology. e-mail: f-elena2008@yandex.ru